REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*
26/09/2007

2. REPORT TYPE
Final Report

3. DATES COVERED *(From - To)*
June 2003 - September 2007

4. TITLE AND SUBTITLE

~~Human and Machine Classification of Active Sonar Echoes~~

5a. CONTRACT NUMBER
N00014-01-G-0460 / 0023

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. Jack McLaughlin
Dr. Scott Philips
Dr. James Pitton

5d. PROJECT NUMBER
Project #398720 / Budget #62-8291

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Applied Physics Laboratory
University of Washington
1013 NE 40th Street
Seattle, WA 98105

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
Code 321US - Undersea Signal Processing Team
875 North Randolph Street, Suite 1425
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This research develops a framework for employing perceptual information from human listening experiments to improve automatic event classification. We focus on the identification of new signal attributes, or features, that are able to predict the human performance observed in formal listening experiments. Using this framework, our newly identified features have the ability to elevate automatic classification performance closer to the level of human listeners.

We develop several new methods for learning a perceptual feature transform from human similarity measures. We also develop a new approach for learning a perceptual distance metric. Our research demonstrates these new methods in the area of active sonar signal processing and confirms anecdotal evidence that human operators are adept in the task of discriminating between active sonar target and clutter echoes. We identify perceptual features and distance metrics using our novel methods. The results show better agreement with human performance than previous approaches.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John Tague, Office of Naval Research
U	U	U	SAR	68	19b. TELEPHONE NUMBER *(Include area code)* 703-696-4399

Perceptually-Driven Signal Analysis for Acoustic Event Classification

Jack McLaughlin and Scott Philips and James Pitton

Applied Physics Laboratory
University of Washington
1013 NE 40th St.
Seattle, WA 98105-6698

Award Number: N00014-01-G-0460, delivery order #23

September 26, 2007

Abstract

In many acoustic signal processing applications human listeners are able to outperform automated processing techniques, particularly in the identification and classification of acoustic events. The research discussed in this paper develops a framework for employing perceptual information from human listening experiments to improve automatic event classification. We focus on the identification of new signal attributes, or features, that are able to predict the human performance observed in formal listening experiments. Using this framework, our newly identified features have the ability to elevate automatic classification performance closer to the level of human listeners.

We develop several new methods for learning a perceptual feature transform from human similarity measures. In addition to providing a more fundamental basis for uncovering perceptual features than previous approaches, these methods also lead to a greater insight into how humans perceive sounds in a dataset. We also develop a new approach for learning a perceptual distance metric. This metric is shown to be applicable to modern kernel-based techniques used in machine learning and provides a connection between the fields of psychoacoustics and machine learning.

Our research demonstrates these new methods in the area of active sonar signal processing. There is anecdotal evidence within the sonar community that human operators are adept in the task of discriminating between active sonar target and clutter echoes. We confirm this ability in a series of formal listening experiments. With the results of these experiments, we then identify perceptual features and distance metrics using our novel methods. The results show better agreement with human performance than previous approaches. While this work demonstrates these methods using perceptual similarity measures from active sonar data, they are applicable to any similarity measure between signals.

Chapter 1

Introduction

1.1 Research Statement

For many aural (acoustic) signal processing tasks, humans are known to perform better than automated classification systems. For such applications, it may be beneficial to identify aspects of the approach used by humans and integrate those aspects into an automatic classification system. In applications such as speech recognition, superior human performance could be the result of high level processing (*e.g.* language models); in other applications, such as identification of transient signals, the key to human performance may lie closer to the periphery, perhaps in the identification and extraction of useful acoustic signal features for classification. Current features used in transient signal classification do not always provide acceptable performance; accordingly, new features are desired that yield the superior classification performance observed in humans. This research focuses on the acoustic feature problem. Specifically, we utilize the results of formal listening experiments to examine short duration transient signals from active sonar systems with the goal of learning new potentially useful feature transforms to aid automatic classification.

In the processing of transient and other nonstationary acoustic signals, features are commonly derived from the statistics of the acoustic signal. These statistical features are calculated from a number of different signal domains, such as, time, frequency, and joint time-frequency. Using a set of training signals, the statistical features with the greatest discrimination power are identified by the researcher and subsequently used to build an automatic classifier. This data-driven approach has yielded numerous features, but these features rarely provide the same classification ability as observed in humans. An alternate approach for identifying new features is a perceptually-driven one, in which features are derived based on their relevance to perception.

In order to scientifically identify what acoustic signal information humans find perceptually important, formal listening experiments are required. These experiments are used to gather numeric measures of how a set of sounds are perceptually organized. For example, in one such experiment humans are asked to listen to pairs of sounds and rate their similarities on some predefined scale. Their similarity judgments reflect an underlying perceptual feature space that humans use when comparing these sounds. Traditional psychoacoustics involves correlating a known set of signal features against the results of these experiments, allowing researchers to access the perceptual relevance of each feature [1]. The limitation of this approach is that it confines the researcher to previously defined signal features. A new approach is desired that can learn the signal features that correspond to observed perceptual results while not requiring *a priori* choices of candidate features.

The goal of the research outlined in this paper is to bridge the work in analyzing perceptual information with modern signal processing techniques for feature identification. The signal processing approach used

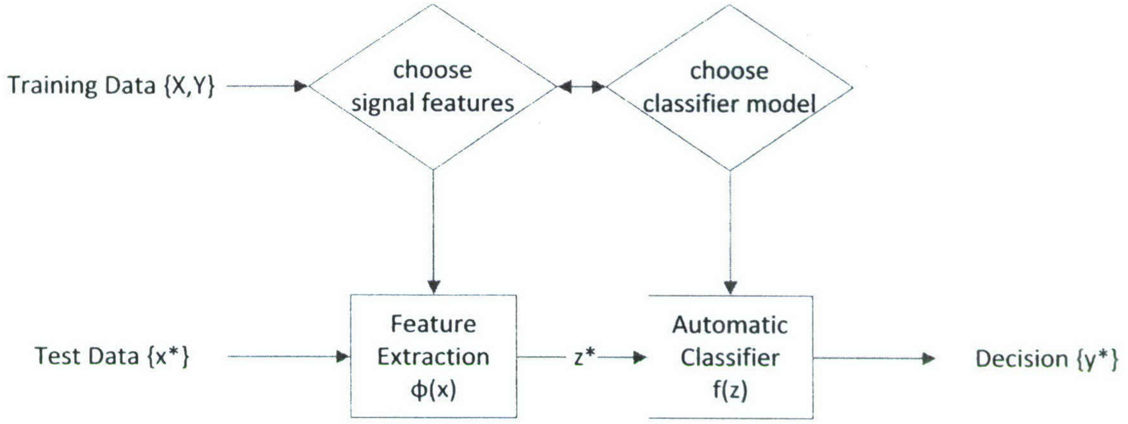


Figure 1.1: Flowchart illustrating the design of an automatic classification system.

by data-driven methods is extended from simple class separation to a regression against the results of a listening experiment. This expands traditional perceptual analysis from simple hypothesis testing to a perceptually-driven "feature discovery" process. While the proposed methods will be applied to the results of a similarity listening experiment using active sonar data, it is applicable to other perceptual domains (*e.g.* visual) and similarity measures.

1.2 Approach

The task of classifying an acoustic signal can be broken into two major stages. First, a number of signal attributes (or features) are extracted from the acoustic signal. This strips the signal of excess information and results in a low-dimensional feature space in which (hopefully) the signal is well described. Next, a class decision is made based on a specific model of this feature space. This model assigns a class label to the acoustic signal based on its location in the feature space. The goal of research in the area of pattern recognition is therefore to identify strong feature spaces and classifier models.

Figure 1.1 shows a flowchart illustrating the design stages in building an automatic classifier. Typically the feature space and classifier model are chosen based on a set of training data. This data is a collection of example signals, $\{X\}$ and associated class labels, $\{Y\}$, that were collected roughly in the same environment that the classifier will operate in. From this data, signal features are identified such that training signals of differing classes are well separated. Once specific features are isolated, a classifier model is then built based on the location of the training data in this feature space. In this research we focus on the first step, identifying an appropriate feature space because, without a descriptive feature space, a reliable class decision cannot be made by any model.

In this paper we examine the perceptual feature space used by human listeners. We include perceptual judgements, $\{P\}$, from human listening experiments in addition to the standard training data, $\{X, Y\}$, when building an automatic classifier. As humans often outperform standard automatic classification systems, this additional information can lead to identification of new and more useful signal features. Even though we cannot directly access the perceptual feature space humans use when judging sounds, listening experiments allow us to gather information about it. Comparative judgements (*e.g.* similarity ratings) from human listeners provide us with insight as to how signals are arranged in their own perceptual space. Using these judgements, this research develops a framework for learning a transformation from a calculable feature space $\mathcal{N}$ to a perceptual feature space $\mathcal{P}$,

$$\mathcal{P} = T\{\mathcal{N}\} \tag{1.1}$$

The following chapter in this paper provides a background of the area of pattern recognition from both a perceptual and statistical perspective. Next, the area of active sonar is introduced. In this chapter we describe three datasets that will be used in subsequent chapters. In chapter 4 a series of listening experiments that we conducted are described. These experiments evaluate human classification ability as well as provide detailed information regarding the preception of active sonar echoes. Chapter 5 and 6 introduce a series of new approaches that utilize the results of these listening experiments to uncover new feature domains and distance metrics. Chapter 7 describes an approach for perceptually weighting a set of training data. This weighting is shown to improve the estimation of the Bayesian likelihood function. Finally, conclusions and future work are discussed in Chapter 8.

Chapter 2

Pattern Recognition

This chapter introduces some fundamental topics that will be covered in depth in this paper. We review current theory in the area of pattern recognition from both a psychoacoustic perspective as well as a statistical perspective. In the area of psychoacoustics we review signal detection theory as well as various perceptual experimental designs. We then provide an overview of statistical pattern recognition, focusing on feature identification and how it relates to classifier design. The research presented in this paper is meant to bridge the gap between the work that has previously been done in these two distinct areas.

2.1 Psychoacoustic Pattern Recognition

Psychoacoustic pattern recognition is an area within psychoacoustics dedicated to understanding how humans perceive differences between sounds. This area is often referred to as the study of musical timbre. Defined as "the subjective attribute of sound which differentiates two or more sounds that have the same loudness, pitch and duration" [2], timbre has been dubbed the psychoacoustician's multidimensional wastebasket category [3]. This characterization has evolved because timbre has been used to describe just about everything regarding a sounds perceptual nature. Psychoacoustic pattern recognition attempts to pin down specific signal attributes that comprise a sound's timbre.

Carefully designing a series of listening experiments is the first challenge when attempting to measure any psychoacoustic information. The following section provides a short introduction into the area of signal detection theory in psychology. It reviews the processes that are thought to be performed when someone makes a perceptual signal classification. We then introduce two experimental designs and discuss the previous work in analyzing the results of these experiments.

2.1.1 Signal Detection Theory

Perceptual classification can be viewed as a decision humans make based on both sensory and cognitive evidence [4]. To follow how this decision is made, we break down a perceptual classification into two independent processes: a sensory process and a decision process. Figure 2.1 illustrates these processes.

First, a sensory process receives the incoming stimuli and converts them from acoustic pressure waves to relevant perceptual cues or sensory evidence. This sensory evidence provides the bases for which a decision can be made. Next, a boundary must be drawn that separates sensory evidence into distinct classes. This boundary is made based on an individual's willingness to identify a sound as part of a particular class. This willingness, or cognitive evidence, can be seen as a subject's understanding of what sensory evidence comprises a sound from a particular class. By combining both sensory and cognitive evidence, a person can identify a sound as a member of a particular class.

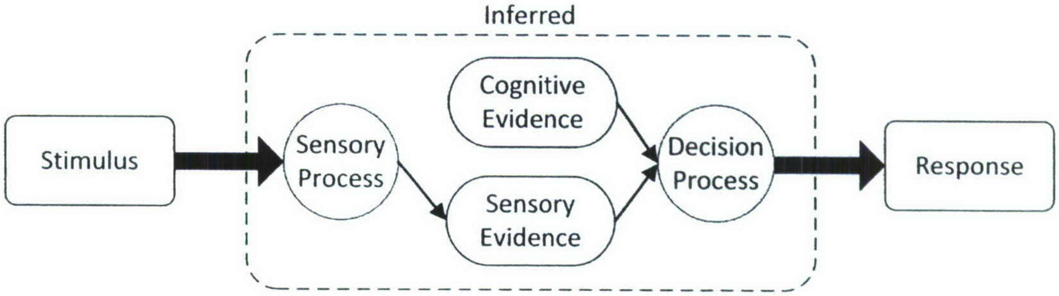


Figure 2.1: Flowchart of the processes that occur in a perceptual classification task.

Each of these stages is a separate process with its own source of variability that needs to be modeled. Sensory variability results from the diversity of sounds within a particular class. Decision variability comes from criterion shifts between subjects in an experiment. A criterion shift is the difference between subjects in their willingness to identify a sound as belonging to a particular class. For the purposes of this research we are interested in gaining an understanding of human sensory evidence, rather than where an individual subject places his decision criterion. Therefore, in the following experiments, we will utilize a variable criterion signal detection model [4–6] in order to isolate human sensory evidence.

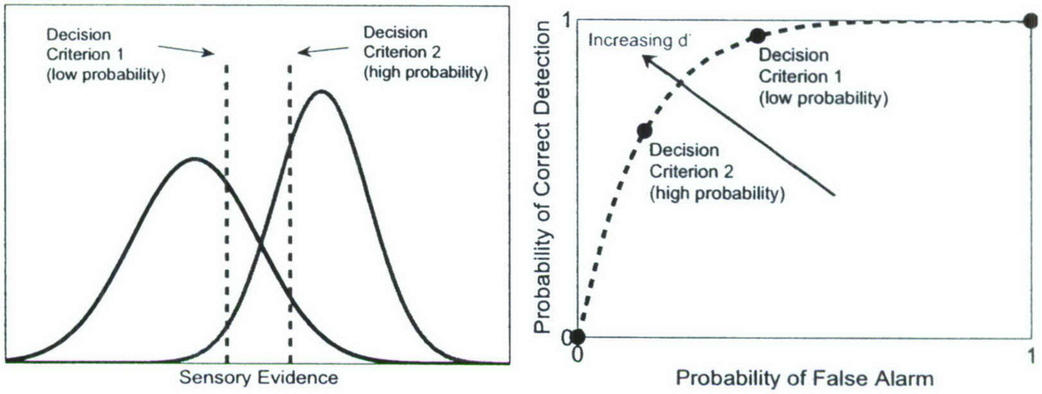
2.1.2 Aural Classification

In order to investigate perceptual classification, we first consider a straightforward classification experiment to assess how well humans can aurally identify sounds of different classes. In addition to providing a baseline performance level, this experiment determines if subjects can perform the classification task at all. For the purpose of this paper we focus on the two-class problem, which we henceforth refer to as target vs. clutter.

This experiment is deceptively simple. It must be carefully designed in order to accommodate for changes in decision criterion between subjects. Therefore, aural classification experiments typically employ a multi-level rating scale as opposed to requiring a subject to definitively identify a sound as a member of a particular class [7]. One common multi-level rating scale is a probability scale, where a subject is asked to rate the probability a sound came from the target class. This probability rating allows the subject to assign a confidence rating to each sound. This rating provides the researcher with multiple decision criteria, one for each probability level.

In order to assess how well a subject is performing, the percentage of correctly identified targets (hit rate) and incorrectly identified clutter (false alarm rate) needs to be calculated. With a multi-level rating system, multiple sets of hit and false alarm rates can be calculated by using different probability thresholds. For example, one threshold treats both low and high probability responses as a rating for a target, and another treats only high probability responses as a rating for target. These thresholds are assumed to lie along some internal perceptual dimension that separates the sensory driven target distribution from the clutter distribution. If we assume both targets and clutter are normally distributed along this perceptual dimension, then we can extrapolate a receiver operating characteristic (ROC) curve from these operating points. This approach is referred to as a binormal model and is illustrated in Figure 2.2. Violations of this assumption often have little practical significance in assessing a subject's classification ability [4]. All perceptual ROC curves generated for this paper were calculated from a maximum likelihood fit to a binormal model using the online software JROCFIT [8].

Using this model, overall sensory performance can be quantified by employing d' ("d prime") sensitivity



(a)

(b)

Figure 2.2: a) Perceptual representation of sounds from two different classes under a binormal assumption; b) interpolated ROC curve from a multi-level rating scale and binormal assumption.

analysis,

$$d' = \frac{\mu_t - \mu_c}{\sqrt{\frac{1}{2}\left(\sigma_t^2 + \sigma_c^2\right)}} \tag{2.1}$$

where μ_t and μ_c are the means and σ_t and σ_c are the standard deviations of the target and clutter distributions respectively. A d' value of zero indicates the target and the clutter distributions are completely overlapping. As d' increases in value from zero, the separation between the target and clutter distributions increases, indicating greater sensitivity. This sensitivity is also apparent in the ROC curve as the curve bows away from the diagonal line (which indicates chance performance), until at extreme values it is along the outer walls of the graph (low false alarm, high hit rates).

2.1.3 Aural Similarity

In order to get a better understanding of how humans perform an aural classification task, more detailed perceptual information regarding the sound's timbre is needed. Most quantitative approaches describing timbre perception use some measure for perceptual distance between sounds [9]. These perceptual distance measures are usually gathered directly from a similarity experiment in which subjects are asked to rate the similarity between pairs of sounds. This type of experiment provides a distance metric describing a subject's underlying timbre space. We wish to quantify this timbre space using physical signal attributes.

As timbre is most likely a multidimensional attribute, most studies in timbre rely on a numerical technique known as multidimensional scaling (MDS) [1, 10–15]. MDS is a nonlinear data analysis technique which takes known (perceptual) distances between data points and identifies a low dimensional Euclidian space that maintains those distances [16, 17]. This projection is accomplished through an iterative process where data points are moved around in a P dimensional space until some goodness of fit criterion is met. In psychoacoustic applications, a version of MDS called nonmetric MDS is most commonly used. In nonmetric MDS the goodness of fit criterion is referred to as "stress" and is defined as

$$S = \sum_{i,j} \left(d(x_i, x_j) - f\left(\delta(x_i, x_j)\right)\right)^2 \tag{2.2}$$

where $d(\cdot,\cdot)$ is the Euclidian distance between two datapoints in the MDS space and $\delta(\cdot,\cdot)$ is the recorded perceptual distance between two points. The function $f(\cdot)$ is a nonlinear scaling function that allows for a nonlinear relationship between perceptual distance and Euclidian distance. While absolute location in the MDS space is arbitrary, the relative locations between sounds reflect the perceptual distance measures.

The space produced by MDS can be viewed as a perceptual feature space representing the acoustic cues subjects use when judging aural similarity [1]. The problem inherent in this approach is that the axes produced by MDS analysis are not labeled. The goal of MDS studies is to identify physical signal features that define these axes, thus quantifying the perceptual signal representation. Most research in this area has focused on identifying these signal features *a priori* and using MDS analysis as a validation tool. Feature validation is accomplished by correlating a hypothesized signal feature for each sound against their corresponding location along each dimension of the MDS space (*e.g.* [10, 12, 14, 15]). This correlation provides a metric defining the degree of perceptual relevance of a hypothesized feature. While a number of perceptually-relevant features have been identified using this methodology, previous studies have not provided an approach for identifying new features from the perceptual similarity data. The approaches outlined in Chapters 5 and 6 introduce a new methodology for identifying new perceptually relevant features from similarity data.

2.2 Statistical Pattern Recognition

Statistical pattern recognition is the process of automatically identifying patterns within raw data based on learning/modeling relationships within that data. In acoustic signal classification, the patterns which are learned are the class labels of the signals (*e.g.* speech recognition). In this section we review previous work in the area of acoustic signal classification, focusing on the feature identification stage.

2.2.1 Features

The first step to automatically classify an acoustic signal is to identify an appropriate feature space. Feature identification is the process of extracting relevant information from an acoustic waveform in order to provide a low dimensional signal description to a classifier model. In many applications researchers simply choose features based on what has worked well in previous applications. These features typically encode information about the shape of the acoustic waveform in some domain, for example, the time domain, the spectral domain or even the time-frequency domain. There has also been interest in identifying application-specific signal features. These features are derived based on a set of training data for a specific classification task. Below we list a few commonly used feature spaces used in previous applications as well as some known methods for identifying data dependent feature spaces.

Data-Independent Feature transforms

A simple way to encode a signal's information is to calculate a series expansion. One common series expansion used in transient signal classification is time envelope moments [18, 19]. These time-series features are defined as

$$\phi_n = \sum_t t^n \frac{v(t)}{\sum_t v(t)} \tag{2.3}$$

where v(t) is the envelope of time series x(t). This envelope can be calculated using the Hilbert transform [20]. These moments provide the center of mass, spread, skew and kurtosis of the signal's time series envelope. Similar to time moments, spectral moments are also used as features in order to encode a signal's spectral shape [18, 21]. These moments are defined as

$$\phi_n = \sum_{\omega} \omega^n \frac{|X(\omega)|^2}{\sum_\omega |X(\omega)|^2} \tag{2.4}$$

Spectral moments provide the signal's spectral center frequency, bandwidth, skew and kurtosis.

Another approach used to capture spectral shape is a cepstral transform. The cepstral transform is defined as

$$c(\nu) = \frac{1}{2\pi} \sum_{\omega} \log |X(\omega)| \, e^{j\nu\omega}. \tag{2.5}$$

In many applications, particularly speech recognition, this transform has been shown to compact spectral energy into a small number of cepstral coefficients [22]. This energy compaction allows the spectrum to be well-described by only a few numbers.

A parametric approach to modeling spectral shape is autoregressive (AR) modeling. This approach models a signal's spectra using the form

$$\hat{X}(\omega) = \frac{1}{1 - \sum_{k=1}^{p} a_k e^{-j\omega k}} \tag{2.6}$$

with AR parameters $\{a_k\}_{k=1}^{p}$. By finding appropriate values for these parameters we can approximate a frequency response $X(\omega)$. Methods for finding the parameters of an AR model include Yule-Walker, Levinson-Durbin, and Burg's algorithm [23].

In applications where a signal is time-varying, it is sometimes useful to identify features from a joint time-frequency signal domain. One specific time frequency representation (TFR) is the spectrogram. This is defined as

$$S(t,\omega) = \left| \sum_{\tau} w(\tau) x(t+\tau) e^{-j\omega\tau} \right|^2 \tag{2.7}$$

where $h(t)$ is a finite duration windowing function centered at t [24]. Features calculated from this domain often consider time and frequency disjointly. One can calculate spectral features as they progress over time or one can calculate time features as they progress over frequency. The most common example of this is in speech recognition where new cepstral coefficients are calculated over a series of time windows (or frames) [22].

Another way to represent a signal with time varying information is in the modulation spectral domain [25]. One definition of this domain is as the Fourier transform across time of the spectrogram,

$$M(\eta,\omega) = \int S(t,\omega) e^{-j\eta t} dt \tag{2.8}$$

where ω is standard acoustic frequency and η is the modulation frequency representing each frequency's time-varying structure. In this domain, spectral features can be identified across either acoustic frequency or modulation frequency.

The above features can be blindly applied to a classification task. The features generally describe a signal and therefore could be used in the task of classifying between different signals. Alternately, there have been a number of methods developed for identifying a set of application specific signal features. These methods uncover features that only describe signal information that is relevant to a particular task and dataset. Below we review some of the time-frequency methods used to identify these types of features.

Data-Dependent Feature transforms

Many data-dependent feature transforms have been developed from within the wavelets community. The previous feature sets discussed have individually considered either frequency features over time or time features over frequency. Wavelet and other orthogonal decomposition methods, however, have the ability to identify features over both time and frequency (scale). The most well known of these methods is the 'best-basis' algorithm [26] (or the related local discriminant basis (LDB) [27]). These methods decompose the time-frequency plane into a large library of 1-D orthogonal basis functions. Commonly these basis functions are derived from an overdetermined wavelet packet and/or cosine packet decomposition [28]. The 'best-basis' algorithm searches through this library for a set of basis functions that minimize an information cost metric (*e.g.* entropy) for a signal (or set of signals for LDB). In effect, this method identifies regions of time and frequency (scale) that are most relevant for describing a training signal(s).

The features found using these wavelet methods are calculated given a specific signal representation (e.g. cosine packet or Haar wavelet decomposition). This approach does not address the problem of which representation is best suited for identifying features. Uncovering a representation that highlights the differences in time-frequency between signal classes can aid in the identification of useful classification features. Recent work by Atlas *et al.* [29, 30] as well as Davy *et al.* [31–33] have considered this problem using the broad class of time-frequency representation's (TFRs) commonly referred to as the Cohen class [24].

Cohen Generalized Time-Frequency Representation The Cohen approach for the design of a generalized TFR arises from the Fourier transform in time t applied to the instantaneous autocorrelation function $R(t,\tau) = x^*(t-\frac{\tau}{2})x(t+\frac{\tau}{2})$. This transform defines the auto-ambiguity function

$$\begin{aligned} A(\theta,\tau) &= \mathcal{F}_{t\to\theta}\{R(t,\tau)\} \\ &= \sum_t R(t,\tau)e^{-j\theta t} \end{aligned} \tag{2.9}$$

where θ and τ represent Doppler and delay respectively. The auto-ambiguity domain defines the characteristic function of the Wigner TFR. Therefore, the Wigner distribution is equal to

$$\begin{aligned} W(t,\omega) &= \mathcal{F}^{-1}_{\theta\to t}\{\mathcal{F}_{\tau\to\omega}\{A(\theta,\tau)\}\} \\ &= \frac{1}{2\pi}\sum_\theta\sum_\tau A(\theta,\tau)\; e^{j\theta t}e^{-j\omega\tau}. \end{aligned} \tag{2.10}$$

From the Wigner distribution we can extrapolate a general class of TFRs by defining a weighting function $h(\theta,\tau)$. This kernel operates multiplicatively upon the auto-ambiguity function providing a generalized TFR given by

$$\begin{aligned} G(t,\omega) &= \mathcal{F}^{-1}_{\theta\to t}\{\mathcal{F}_{\tau\to\omega}\{h(\theta,\tau)A(\theta,\tau)\}\} \\ &= \frac{1}{2\pi}\sum_\theta\sum_\tau h(\theta,\tau)A(\theta,\tau)\; e^{j\theta t}e^{-j\omega\tau}. \end{aligned} \tag{2.11}$$

A weighting function of $h(\theta,\tau) = 1$ corresponds to the Wigner distribution. Any other nonzero function will induce an implicit smoothing of the Wigner distribution in time and/or frequency. Any version of $G(t,\omega)$ is therefore a smoothed versions of $W(t,\omega)$. The power of this representation is that *all* quadratic TFRs can be obtained from $W(t,\omega)$ by the application of the appropriate weighting function [24, 30]. For example, in this framework the spectrogram, $S(t,\omega)$, is defined by the weighting function $h(\theta,\tau) = \sum_u w^*(u-\frac{\tau}{2})w(u+\frac{\tau}{2})e^{j\theta u}$ [24].

Optimizing the Cohen Kernel for Classification The development of TFRs is often motivated by the expressed goal of accurately describing a signal in both time and frequency. While important for visualization, this goal may not be optimal for signal classification. Methods have proposed leveraging TFR design with the expressed goal of classification [29–32]. These methods are used to provide a TFR more suited to highlight differences between signals of different classes.

To illustrate this approach, consider a two class supervised classification problem in which there exists n_c examples in the training set for each class. Let $x_i^{(c)}$ represent the i^{th} training example from the c^{th} class. Define the class average TFR as

$$\overline{\mathbf{G}}^{(c)} = \frac{1}{n_c} \sum_{i=1}^{n_c} \mathbf{G}_{x_i^{(c)}} \tag{2.12}$$

where $\mathbf{G}_{x_i^{(c)}}$ is the matrix representation of the generalized TFR derived from $x_i^{(c)}$. Also define the Kolmogorov distance between generalized TFRs as

$$d_k \left(\mathbf{G}_{x_1}, \mathbf{G}_{x_2}\right) = \|\mathbf{G}_{x_1} - \mathbf{G}_{x_2}\|_F^2 \tag{2.13}$$

where $\| \cdot \|_F^2$ is the Frobenius norm of a matrix.

Using this notation, a class dependent function h_{CD} is found that optimizes a specific classification criteria. In [29, 30] a class dependent function is defined by

$$h_{CD} = \underset{h}{\operatorname{argmax}} \left\| \overline{\mathbf{G}}^{(1)} - \overline{\mathbf{G}}^{(2)} \right\|_F^2 \tag{2.14}$$

This criterion defines an h_{CD} that maximizes the mean distance between signals from different classes. A variant of this criterion was proposed in [31] (and similarly in [30]) in which the mean distance between classes was maximized subject to within-class variance. This criterion is

$$h_{CD} = \underset{h}{\operatorname{argmax}} \frac{\overline{d_k} \left(\mathbf{G}_{x^{(1)}}, \overline{\mathbf{G}}^{(2)}\right) + \overline{d_k} \left(\mathbf{G}_{x^{(2)}}, \overline{\mathbf{G}}^{(1)}\right)}{\overline{d_k} \left(\mathbf{G}_{x^{(1)}}, \overline{\mathbf{G}}^{(1)}\right) + \overline{d_k} \left(\mathbf{G}_{x^{(2)}}, \overline{\mathbf{G}}^{(2)}\right)} \tag{2.15}$$

where $\overline{d_k} \left(\mathbf{G}_{x^{(i)}}, \overline{\mathbf{G}}^{(j)}\right)$ denotes the average Kolmogorov distance of TFRs in class i to the average TFR from class j. This second maximization is a Fisher-like contrast criterion.

In practice both equation 2.14 and 2.15 lead to a very high-dimensional functional optimization. Without any assumptions on the structure of h_{CD} this would be an intractable task. Therefore certain constraints must be imposed. In [30], Atlas *et al.* restricts the weighting function to have only a finite number of nonzero values. In this approach only the points in θ and τ which provide the best separation in TF are used in ϕ_{CD}. These point are identified by rank ordering individual all of kernel points according to class separation.

In [31, 32], Davy *et al.* restricts the weighting function to a functional form defined by a limited number of parameters. Noticing that the TFR is real-valued if and only if the corresponding kernel is real and symmetric, an appropriate kernel shape would be radially Gaussian [34]. In polar coordinates ($\rho^2 = \theta^2 + \tau^2$ and $tan(\psi) = \theta/\tau$) define the radial Gaussian kernel as

$$\phi(\rho, \psi) = e^{-\frac{\rho^2}{2\sigma^2(\psi)}} \tag{2.16}$$

The "contour function" $\sigma(\psi)$ determines the final shape of the kernel. It specifies the bandwidth of the Gaussian shape for a given angle ψ. To keep the kernel symmetric σ is defined via a truncated Fourier series:

$$\sigma(\psi) = a_0 + \sum_{i=1}^{p} a_i \cos(2i\psi) + b_i \sin(2i\psi) \tag{2.17}$$

where a and b are the Fourier coefficients. With this restriction the optimization of ϕ_{CD} only has $2p+1$ parameters to estimate.

2.2.2 Kernel functions

Kernel functions are an alternative method for describing the characteristics of a signal. Instead of directly calculating signal features, kernels simply use relational measures between signals. The most common form of a kernel is a Mercer kernel, which is defined as

$$K(x_i, x_j) =< \phi(x_i), \phi(x_j) > \tag{2.18}$$

where $\phi(\cdot)$ is a mapping from input space to a feature space [35]. The power of this approach is that the underlying feature space does not need to be defined explicitly; only the function that measures the relation between signals $K(x_i, x_j)$ is required. These kernel functions are used in modern classifier models such as the support vector machine (SVM) [36] and the relevance vector machine (RVM) [37, 38].

More recent classifier models have been able to relax the strict Mercer kernel definition. New classifier models do not require the relational measure to be of inner product form. Modern kernel methods can handle any distance or similarity metric between signals. Classifier models that make use of these relaxed kernel are Gupta and Cazzanti's similarity-based classifier [39] and the potential support vector machine (pSVM) [40, 41].

Numerous kernel functions have been proposed for automatic classification and regression. Smola and Schölkopf review a number of the more successful kernel functions in [42]. Three examples of these kernels are polynomial kernels

$$K(x_i, x_j) = (< x_i, x_j > +c)^p\,, \tag{2.19}$$

hyperbolic tangent kernels

$$K(x_i, x_j) = \tanh(\kappa < x_i, x_j > +\Theta), \tag{2.20}$$

and most commonly the radial basis kernel

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}. \tag{2.21}$$

These functions have been successfully demonstrated in previous applications and therefore are often blindly applied to new applications without a strong rationale.

Recent work has moved beyond these basic kernel choices. Methods for the identification of data-dependent kernels have been introduced. In 2001, Cristianini *et al.* [43] introduced the concept of kernel-target alignment. They define kernel alignment as

$$A = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \tag{2.22}$$

where $\langle K_1, K_2 \rangle_F = \sum_{i,j} K_1(x_i, x_j) K_2(x_i, x_j)$ is an inner product between Gram matrices for some training data $\{x_i | i = 1, ..., N\}$. The kernel alignment is used to measure the agreement between two kernel functions on a given dataset with 1 meaning perfectly aligned and 0 meaning not aligned at all.

An appropriate choice of kernel function for a given dataset can be determined by comparing the alignment of various kernels against an ideal kernel for a specific task. For example, the ideal kernel function for the two-class problem is $K_I(x_i, x_j) = y_i \cdot y_j$. Cristianini *et al.* propose that this can be used not only to identify which kernels are most useful, but also to combine kernels in order to identify improved kernels for a specific task.

Other studies have attempted to learn a kernel for the application of semi-supervised learning in which kernel functions are learned from a mixture of labeled and unlabeled data. Examples of these include Chapelle *et al.* [44] and Zhu *et al.* [45]. Methods in this area include using a measure of kernel alignment to optimize the parameters of a kernel model as well as using semi-definite programming for learning a non-parametric kernel function.

Chapter 3

Application and Data

In this chapter we will introduce the application used to demonstrate our proposed methodologies, active sonar. First we will provide a brief overview of active sonar. Afterwards, we will discuss three types of data that have been gathered and specific datasets that are used in this paper.

3.1 Sonar Systems

Sonar (SOund Navigation And Ranging) is a term used to describe systems that extract information about an underwater environment using the propagation of acoustic waves. Broadly, sonar systems can be divided into two categories: passive sonar and active sonar. Passive sonar systems require no sound source of their own. They simply listen to sounds already present in the environment. In contrast, active sonar systems rely on a self generated sound source. Inferences can be made based on how the generated sound interacts with the surrounding environment. The examples shown in this paper deal with data generated from active sonar systems.

A typical active sonar system consists of a source and a receiver. A sound is generated at the source which then propagates throughout the environment. The sound proceeds to scatter off of any nearby objects and that scatter is then recorded at the receiver. An idealized cartoon representation of this operation is shown in Figure 3.1. This figure illustrates the operation of a horizontal line array, where the source is generated at the front of the line array and a series of receivers trail behind.

The most common applications of active sonar systems are in the detection, localization and classification of underwater objects. While current technologies have shown considerable promise in detection and localization, automatic classification of target objects remains a challenging problem. In littoral environments sonar systems are often flooded with echoes from numerous objects in the surrounding water. Current automatic classifiers, while able to distinguish some of the target objects from clutter, tend to be overly sensitive, often mistaking clutter for targets. The techniques developed in this paper will be demonstrated in this area of sonar classification with the goal of using psychoacoustic information to improve the automatic classification of sonar echoes.

3.2 Data

Three active source sonar datasets will be used to validate the methodology outlined in this paper. Below is a description of each of these datasets.

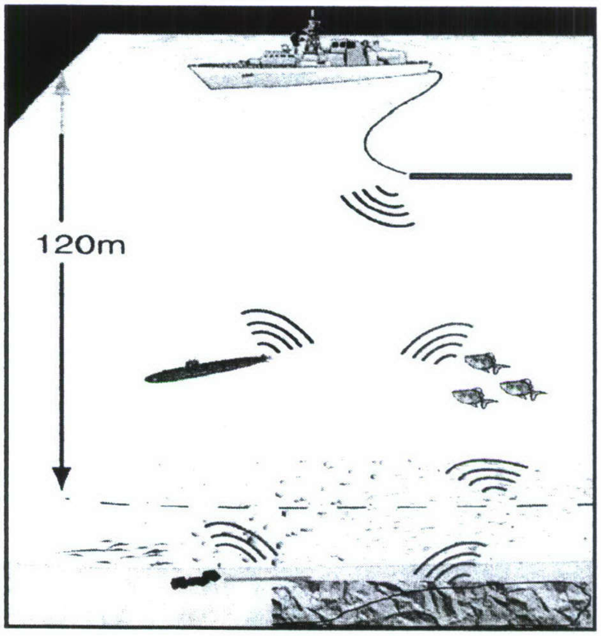


Figure 3.1: Cartoon illustrating the operation of an active sonar system (Source: Defence Research & Development Canada).

3.2.1 U.S. Navy Impulsive-Source Data

The U.S. Navy data is a set of sonar echoes using an impulsive-source active sonar system. All detected echoes have been isolated into short sound files and are labeled as either a true target or false target (clutter) based on the known location of all target objects. The "hard-case clutter" subset of the data is a collection of false target detections that represents all clutter echoes that were misclassified by an automatic classifier. This clutter subset, along with all detected true target echoes comprise the U.S. Navy dataset of sonar echoes. We use these echoes in a series of listening experiment in order to identify perceptual features from the data that will be analyzed throughout this paper. For insight on this dataset, spectrograms of two hard-case clutter echoes are shown in Figure 3.2.

3.2.2 Boundary 2004 Impulsive-Source Data

Boundary 2004 is another dataset of impulsive-source active sonar echoes that was collected off the coast of Sicily in the summer of 2004. Detected echoes have been isolated into short sound files and labeled as true targets and false targets (clutter) based on known locations of the targets. Targets consist of underwater metallic objects such as surface ships, oil platforms, oil pipelines and ship wrecks. Clutter objects consist of any non-target echo. Another "hard-case clutter" subset was identified in order to remove any easily identified clutter examples from the dataset (*e.g.* artificial spikes in energy that passed the detection algorithm). The targets and "hard-case clutter" echoes comprise the second dataset that will be used in the listening experiment of Chapter 4. Spectrograms of a target and clutter echo from this dataset are shown in Figure 3.3

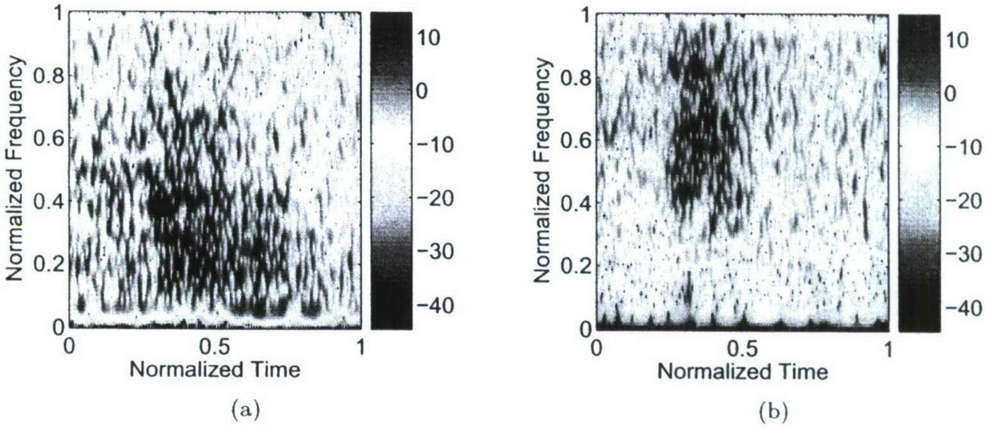


Figure 3.2: Spectrograms of two hard-case clutter echoes from the U.S. Navy dataset.

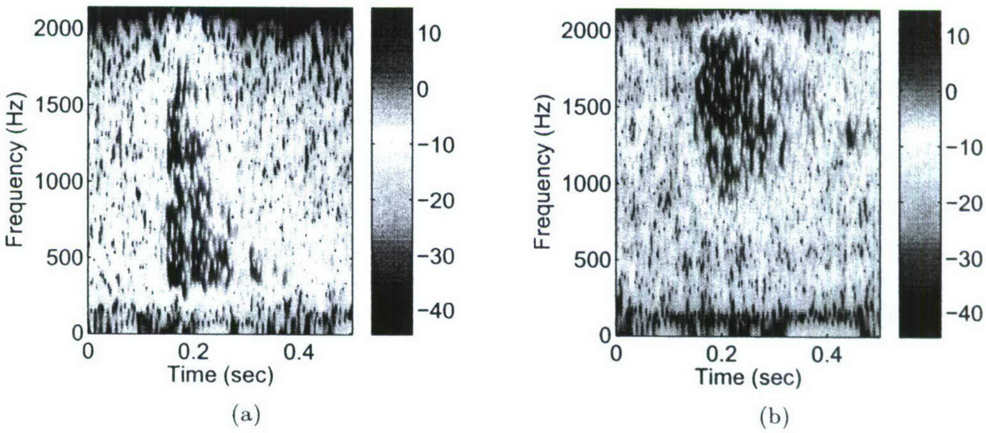


Figure 3.3: Spectrograms of a) target and b) clutter echoes from the Boundary 2004 dataset.

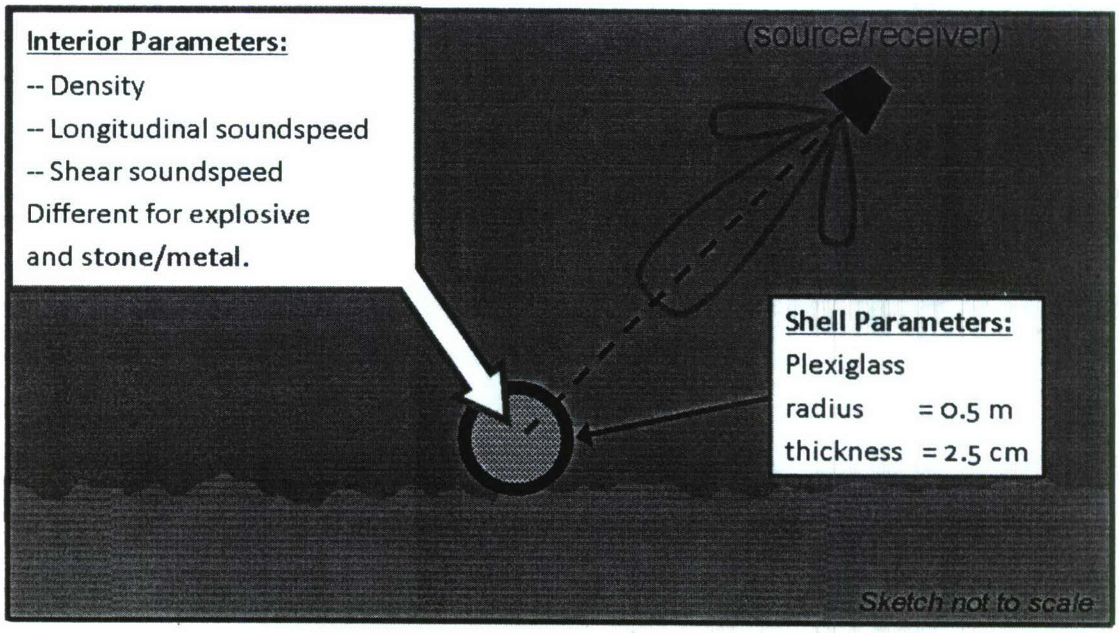


Figure 3.4: Experimental setup for modeling the acoustic response of spheres with varying physical parameters

3.2.3 Modeled Mine Data

The Modeled Mine dataset is a set of numerically modeled sonar echoes from undersea spheres with varying physical structure [46]. The modeled spheres were all 0.5 meter in diameter with a varying internal material contained within a plexiglass shell. These shells were "placed" on a sandy sea floor and a sonar source/receiver was "placed" 50 meters away. A transmitted broadband source was then simulated and the scatter from the sphere was recorded at the receiver. Figure 3.4 illustrates this experimental setup.

The sound properties of the material inside the plexiglass shell can be varied to simulate various filling materials. The main properties of interest are the internal propagation speed of the longitudinal and shear sound waves. These properties are notably different for explosive-filled spheres and stone (and/or metal) filled spheres [47], indicating that they could be used to distinguish between the echoes from these different materials. The location in this parameter space of various explosives and stones/metals is shown in Figure 3.5.

Figure 3.6 presents spectrograms of two of these modeled echoes. Figure 3.6(a) shows a spectrogram of an echo that was modeled from an explosive-filled object and Figure 3.6(b) shows a spectrogram of an echo from a stone-filled object. We can immediately see that there is a difference in resonance structure between these two echoes. If we could identify a signal feature(s) from these echoes that predicted longitudinal and/or shear sound speed then we would be able to distinguish between these different types of underwater objects.

In order to create a larger set of echoes, we fit a Gaussian distribution in the parameter space to the location of the explosives and another distribution to the location of the stones/explosives. A dataset of 25 target echoes and 25 clutter echoes is then drawn from the explosive and stone/metal distributions respectively. The parameter space of these data points is shown in Figure 3.7(a). Figure 3.7(b) shows a matrix of distances between all 50 echoes in the dataset. Using this data we will seek to identify signal features from the echoes that describe this parameter space. Identifying this parameter space is analogous to identifying perceptual features, but instead we are looking signal features that describe the physics of

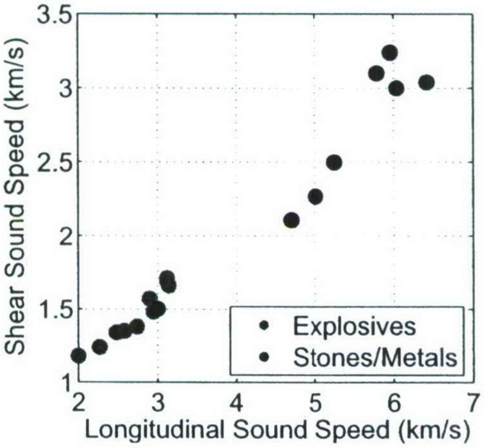


Figure 3.5: Location of different types of explosives and stones/metals in a physical parameter space.

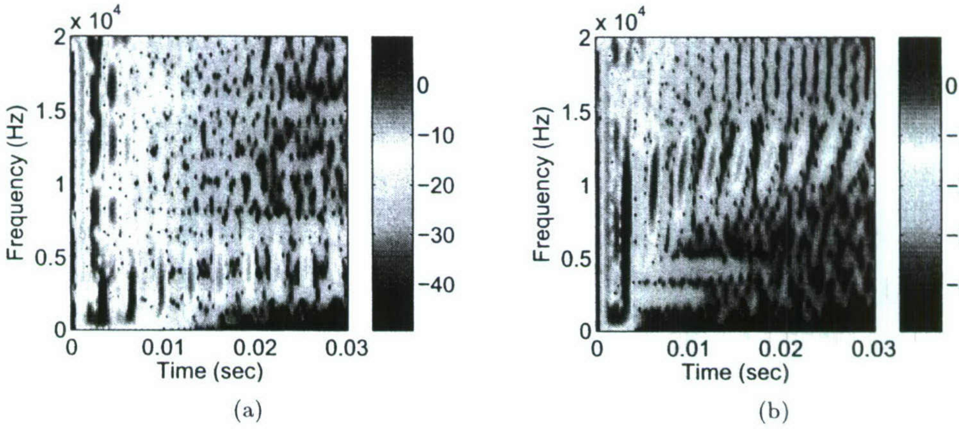
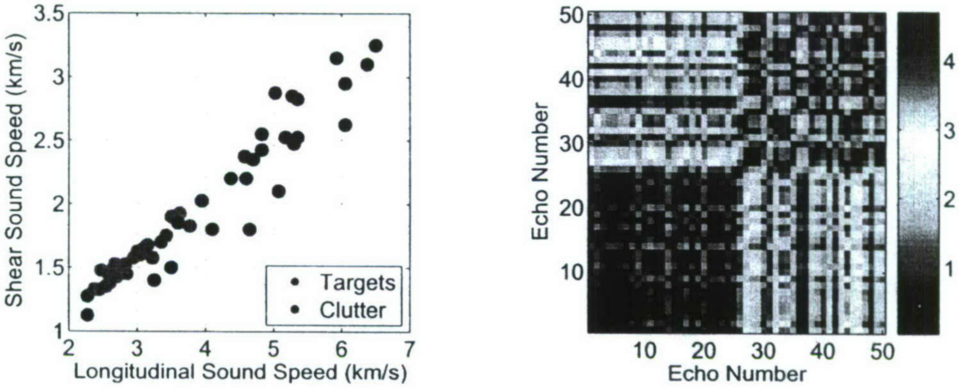


Figure 3.6: Spectrograms of modeled echoes received from a) explosive and b) stone objects.



(a) (b)

Figure 3.7: a) Location of modeled target and clutter objects in physical parameter space and b) distance matrix between those objects (Targets: #1-25, Clutter: #26-50)

the object instead of the perceptual nature of the object.

Chapter 4

Listening Experiments

When attempting to improve automatic classification with aural perceptual information, a series of formal listening experiments must be conducted. This chapter presents the results of a series of experiments that were conducted using trained impulsive-source sonar operators and naïve listeners. The first set of experiments are classification experiments meant to determine how well humans can distinguish target from clutter echoes using aural cues alone. While this ability has been anecdotally reported, our study determines the true performance level of a set of sonar operators and non-expert listeners. Next, the results of a set of similarity experiments are reported. These experiments provide more detailed information regarding how humans perceptually organize these sounds.

4.1 Classification Experiments

In this section, we present the results of three listening experiments using trained impulsive-source sonar operators and naïve listeners to identify targets in a cluttered environment. In the first experiment, we evaluated sonar operator performance in target/clutter discrimination on signals from an impulsive-source active sonar system. The operators' ability to distinguish targets from clutter was used as a standard for comparison in subsequent experiments. In the second experiment, we evaluated the baseline performance of naïve listeners in the same task as the first experiment. In the third experiment, the naïve listeners were given feedback on their decisions, which enabled them to learn to distinguish between targets and clutter over the course of the experiment, bringing their performance closer to the level of experienced sonar operators.

4.1.1 Experiment 1: Evaluation of Sonar Operators

The objective of this experiment was to determine if sonar operators can distinguish targets from environmental clutter using aural cues alone and, if so, determine how well they can perform this task. The resulting level of performance is later used as a baseline to compare the performance of naïve listeners and determine whether they can adequately learn the task.

Experimental Setup

Task: A single sound was played and the subject was asked to respond with one of three probability levels: "none," for no target present; "low," for low confidence of target present; or "high," for high confidence of target present.

Table 4.1: d' values for sonar operators

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**
d'	2.06	1.38	1.63	1.21	0.86	0.38	1.27	1.72	0	-0.11	0.76	1.36	1.74	0.6	1.21

	Ave	**Top 4**
d'	1.07	1.53

Subjects: Fifteen sonar operators from the Patuxent River Naval Air Station participated in Experiment 1. Subjects in this experiment reported varying degrees of experience as sonar operators and familiarity with the task of sonar classification.

Stimuli: The stimuli consisted of 180 impulsive-source sonar signals from the U.S. Navy dataset. The signals contained 90 true targets and 90 "hard-case clutter" echoes that were randomly drawn from the complete dataset.

Presentation: The original .wav files were transferred from a laptop computer running Windows 2000 through a Firewire (IEEE 1394) digital interface to a MOTU 868 audio processor (24-bit, 96 kHz). The MOTU performed a D/A conversion before being recorded and re-digitized at 44.1 kHz onto a digital audio tape (DAT). From the DAT recordings, audio compact discs were created. Sounds were presented to sonar operators over Sony MDR-V600 headphones and played from a Panasonic portable CD player. Subjects were allowed to replay sounds as needed. No feedback was provided to the operators; that is, the subjects were not informed as to whether or not their decision was correct.

Results and Analysis

Figure 4.1(a) illustrates ROC curves for each subject; Figure 4.1(b) shows average ROC curves with confidence intervals for all subjects averaged together, and the top 4 subjects. These ROC curves were computed according to the method described in section 2.1.2. Table 4.1 summarizes the performance of all of the sonar operators in terms of d' values. As is evident from the ROC curves and the table, a large variation in ability was observed. All but two subjects performed above the level of chance, with the average performance well above chance. These d' values were determined to be significantly greater than zero according to a t-test at a level of one percent.

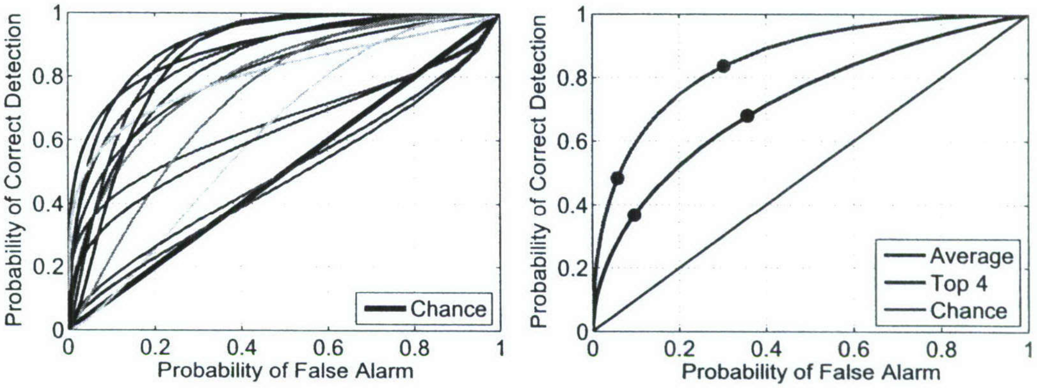
Consistent with the hit and false alarm analysis, the four most sensitive listeners were Subjects 1, 3, 8, 13. The d' value for these subjects was 1.526. These four listeners were the most likely to detect the target and discriminate it from the clutter. Accordingly, these four subjects will be used as a point of comparison to provide a baseline level of performance against which the naïve listeners in the following experiments will be compared.

4.1.2 Experiment 2: Evaluation of Naïve Listeners

The objective of this experiment was to compare the discriminative ability of a set of naïve listeners to that of the experienced sonar operators under the same experimental conditions.

Experimental Setup

Task: Same as Experiment 1



(a)

(b)

Figure 4.1: a) Individual ROC curves for each sonar operator; b) average and top 4 ROC curves with operating points for sonar operators.

Table 4.2: d' values for naïve listeners in experiment 2

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
d'	1.36	1.19	1.11	-0.28	0.74	0.94	1.08	1.28

	Ave	**Top 4**
d'	0.77	1.19

Subjects: Eight naïve listeners were recruited from the Applied Physics Lab at the University of Washington. None of the subjects had operational experience in sonar systems. Their understanding of the sonar classification problem ranged from truly naïve to detailed technical understanding of sonar signal processing.

Stimuli: Same as Experiment 1

Presentation: The naïve listeners were presented the .wav files directly from a computer via a MATLAB$^{\text{TM}}$ script. The sounds were presented in random order through an M-audio A/D board over Sennheiser HD 280 Pro headphones. Subjects were allowed to replay sounds as needed.

Results and Analysis

Figure 4.2 gives ROC curves for each subject as well as average ROC curves for all subjects and the top 4 performing subjects. Table 4.2 summarizes the performance of the subjects in the experiment as before. As in experiment 1, a large variation in ability was observed. Most subjects performed above chance, though not at the high-performance level of the sonar operators. Overall, the d' values were still significantly greater than zero at a level of one percent. The top 4 naïve subjects performed comparable to the average of the sonar operators, but not at the level of the top 4 sonar operators.

Two non-smooth ROC curves can be seen in Figure 4.2(a). Such curves result from a subject not using all decision levels. One subject never called clutter a "high probability detection," resulting in a false alarm rate of zero for the high-threshold decision point. A second subject did not use one of the rating categories altogether.

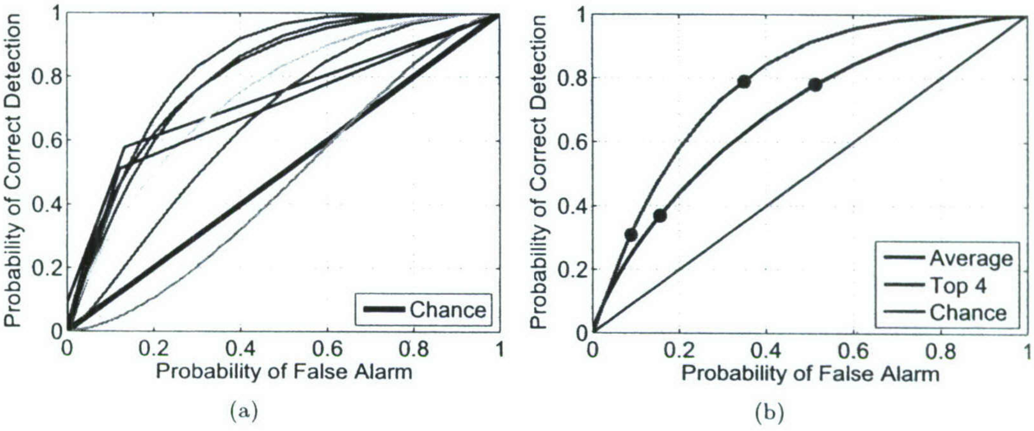


Figure 4.2: a) Individual ROC curves for each naïve listener in experiment 2; b) average and top 4 ROC curves with operating points for naïve listeners in experiment 2.

4.1.3 Experiment 3: Training of Naïve Listeners

The objective of this experiment was to provide training to the naïve listeners to determine whether they could achieve a level of performance comparable to that of the expert sonar operators. Training was in the form of feedback following each trial. The feedback enabled the subject to learn the differences between targets and clutter through a process of trial and error.

Experimental Setup

Task: Same as Experiment 1

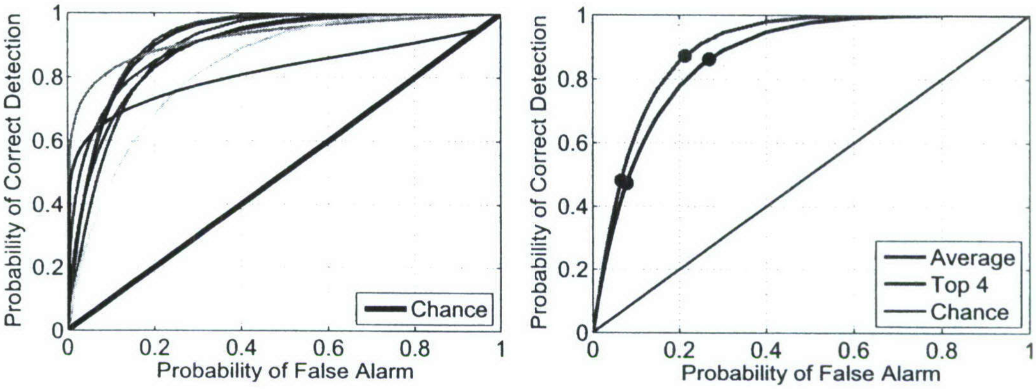
Subjects: Same as Experiment 2

Stimuli: An additional 58 signals were added to the dataset used in experiments 1 and 2, to allow for more training time. This increased the set to 119 true targets and 119 "hard-case clutter" echoes.

Presentation: Same as Experiment 2, except feedback was also given to the subject after each decision, informing them as to whether their response was correct or not.

Results and Analysis

Figure 4.3 and Table 4.3 illustrate subject performance in terms of ROC curves and d' values. The first 50 responses were discarded to exclude learning effects from the analysis. In this experiment, all subjects performed well above chance. There was consistent ability across subjects after allowing time for learning. Feedback clearly improved the classification level, both in higher correct detection rate and lower false alarm rate. The average performance of the naïve listeners with feedback was comparable to that of the top 4 sonar operators.



(a)

(b)

Figure 4.3: a) Individual ROC curves for each naïve listener in experiment 3; b) average and top 4 ROC curves with operating points for naïve listeners in experiment 3.

Table 4.3: d' values for naïve listeners in experiment 3

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
d'	1.72	1.44	1.69	3.26	1.66	1.21	1.39	1.9

	Ave	**Top 4**
d'	1.34	2.08

4.2 Similarity Experiments

In this section, we present the results of two aural similarity experiments using U.S. Navy and Boundary 2004 sonar data sets. These empirical similarity measures provide detailed information on the perceptual distribution of these sounds.

4.2.1 Experiment 4: U.S. Navy Similarity

Experimental Setup

Task: Each subject was presented two sounds in succession. The subject was then asked to rate how similar the sounds were on a scale of one to five (one being very similar and five being very different). After receiving instructions the subjects were given five "practice" trials in order to get a sense of the task. These practice trials were not included in the results or analysis.

Subjects: Same as Experiment 2

Stimuli: As a rule of thumb, the number of stimuli required in multidimensional scaling (MDS) experiments is at least $n = \frac{40k}{m} + 1$, where n is the number of stimuli, k is the number of expected MDS dimensions and m is the number of subjects that rate each pair [16]. In order to decrease the load on the subject while also allowing for the possibility of up to 5 dimensions, a set of 50 targets and 50 clutter echoes were randomly chosen from the U.S. Navy dataset. Even with only 100 stimuli, the number of pair-wise combinations to be judged by each subject would have been 4950. In order to accommodate this

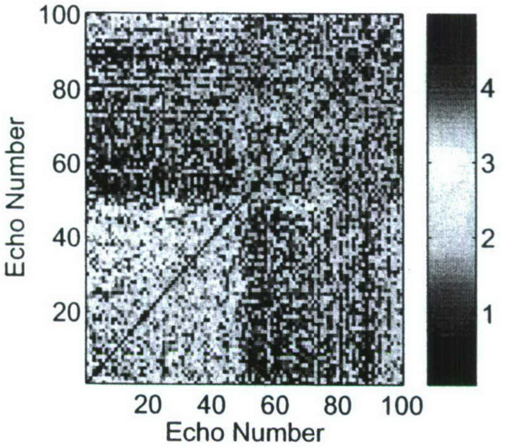


Figure 4.4: Aural similarity matrix for both targets (echoes 1-50) and clutter (echoes 51-100) from the U.S. Navy dataset. Note that only upper right quadrant corresponding to clutter-clutter pairings will be used in the analysis shown in this paper.

large number, the set of pairings was randomly divided into four subsets with each subset rated by two subjects [48].

Presentation: Subjects were presented .wav files directly from a computer via a MATLAB$^{\text{TM}}$ script. The stimulus pairs were presented in random order through an M-audio A/D board over Sennheiser HD 280 Pro headphones. Subjects were allowed to replay stimulus pairs as desired.

Results

Figure 4.4 shows the perceptual similarity matrix δ that results from the subjects' similarity responses. Each entry in the matrix represents one stimulus pair. The matrix is exactly symmetric as only one ordering of the stimuli was used and the resulting data were reflected about the diagonal. Each stimulus pair was presented to two subjects, each of whom rated the similarity between 1 and 5, one being very similar and 5 being very different. The two subjects' responses were averaged and entered into the matrix. Note that the lower left quadrant of the matrix, corresponding to target-target stimulus pairs, has substantially lower values than the rest of the matrix, indicating more similar sounds. The values in the lower right quadrant (target-clutter pairs) are much higher, indicating that targets and clutter were usually judged more dissimilar. The upper right quadrant (clutter-clutter) has a wide range of values compared to the other two quadrants, which suggests that the clutter examples are not a single stimulus class, but rather span a wide range of stimulus types that differ from the target class. This similarity matrix is used in later analysis to identify relevant perceptual features.

4.2.2 Experiment 5: Boundary Similarity

Experimental Setup

Task: Each subject was presented two sounds in succession. The subject was then asked to rate how similar the sounds were on a scale of one to ten (one being very similar and ten being very different). The

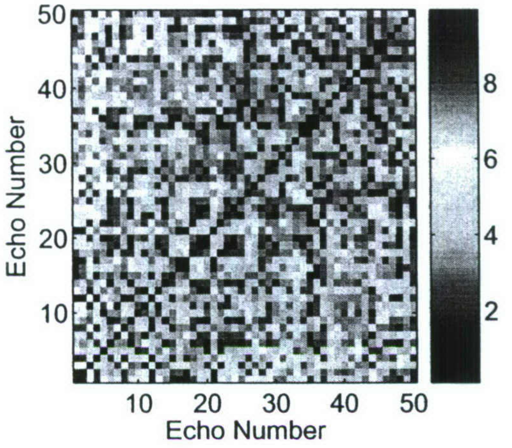


Figure 4.5: Aural similarity matrix for both targets (echoes 1-25) and clutter (echoes 26-50) from the Boundary 2004 dataset.

range was increased from the previous similarity experiment as to give the subjects more freedom with their responses.

Subjects: Eight subjects were recruited from the department of electrical engineering at University of Washington. None of the subjects had operational experience in sonar systems.

Stimuli: All 50 echoes from the Boundary 2004 dataset were used. This set included 25 target echoes and 25 clutter. With 50 stimuli, the number of pair-wise combinations to be judged is 1225. In order to accommodate this large number, the set of pairings was randomly divided into two subsets with each subset rated by four subjects [48].

Presentation: Subjects were presented .wav files directly from a computer via a Matlab script. The stimulus pairs were presented in random order through an M-audio A/D board over Sennheiser HD 280 Pro headphones. Subjects were allowed to replay stimulus pairs as desired.

Results

The results of this similarity experiment are shown in Figure 4.5. Each stimulus pair was presented to four subjects, each of whom rated the similarity between 1 and 10. The four subjects' responses were then averaged and entered into the matrix. As before, the similarity matrix is exactly symmetric as the results were reflected about the diagonal. The lower left quadrant (target-target pairings) again has relatively lower values while the the lower right quadrant (target-clutter pairings) has higher values. While this distinction is not as noticeable as in the previous experiment, there still seems to be some class separation. The increased variability seen in the target-target pairings could be due to the fact that the type of targets in the Boundary 2004 data are not as uniform as in the U.S. Navy data. This similarity matrix will also be used in later analysis to identify new perceptual features.

4.3 Discussion

While there has been much anecdotal evidence, these experiments validate the conventional wisdom that sonar operators are capable of discriminating between impulsive-source active sonar target and clutter echoes solely on the basis of aural cues (listening). This ability indicates there is benefit to further investigation into how subjects perform this task. In addition, it was shown that naïve listeners can be trained to perform the discrimination task to a level comparable to that of sonar operators. The "trained" naïve listeners thus form a subject pool that can be used in place of operators in the more involved similarity listening experiments. Finally, results collected in the similarity experiment also show a strong difference between signals of different classes. This experiment contained more information regarding the distributions of the targets and clutter relative to one another. In the following chapters we will use this additional perceptual information to identify signal attributes that are being used by the subjects.

Chapter 5

Feature Identification

This chapter introduces a new technique for learning perceptually relevant acoustic features found from listening experiment data. The first section describes the standard approach for analysis and feature identification using similarity data. Next, an alternate approach is proposed in which optimal features are found by maximizing their fit to the listening experiment results. These feature identification methods are then demonstrated on the U.S. Navy, Boundary 2004 and Modeled Mine datasets.

5.1 Standard MDS Analysis

Multidimensional scaling (MDS) is a technique used to identify a low-dimensional space in which distance between data points reflects their known degree of similarity [16, 17]. In psychoacoustics, the space found by MDS can be interpreted as a perceptual feature space in which the dimensions represent the acoustic cues that are used in judging similarity. Therefore, the goal of perceptual feature identification is to identify what numeric features best represent each of the dimensions of the MDS space.

In previous studies [10, 12, 14, 15], a number of hypothesized signal features, $\phi_i(\cdot)$, are first calculated for each sound. Each of these features is then correlated with their corresponding location along each dimension of the MDS analysis. This correlation is done via the inner product,

$$\rho_{i,k} = < \phi_i(\mathbf{x}), c_k(\mathbf{x}) > \tag{5.1}$$

where $c_k(\cdot)$ represent the MDS values for along dimension k and $\rho_{i,k}$ represent the correlation between feature i and MDS dimension k. This correlation provides a metric defining the degree of perceptual relevance for each of our hypothesized features. Perceptual features are then identified according to which features provide the highest correlation to each dimension.

Following this approach, we utilize a set of 33 candidate features to correlate against our MDS results. Some of features are standard amplitude and shape statistics calculated from the time series and the spectrum (*e.g.* mean, standard deviation, skewness, kurtosis). Others were perceptually-motivated features found from previous perceptual studies [1, 14, 49]. These features include time-frequency features such as subband correlation, subband rise time, and spectral flux, as well as other classes of features. A complete list of these features along with definitions can be found in Appendix A.

This approach is completely dependent upon the initial choice of features to correlate. It also provides no avenue for uncovering new features using the perceptual data. Thus a new, data driven, approach is desired that can learn the currently unknown signal features that account for perceptual similarity while not requiring an *a priori* choice of candidate features. The following section illustrates an approach for achieving this goal.

5.2 MDS Fitting

In order to identify new signal features that account for perceptual similarity, we follow a systematic approach in which perceptually relevant features are learned by optimizing over a broad class of signal features $\phi_h(\cdot)$. This approach utilizes an infinite set of candidate signal features defined by continuous parameter(s) h as opposed to a finite collection of hypothesized features. By finding the $\hat{h}$ that provides the best fit between the feature vector $\phi_h(\mathbf{x})$ and the MDS data, we can uncover novel features that are relevant to perception. To find this $\hat{h}$, we explicitly maximize the magnitude of the correlation of $g_h(\mathbf{x})$ with the k^{th} MDS dimension according to

$$\hat{h}_k = \operatorname*{argmax}_h \frac{|< \phi_h(\mathbf{x}), c_k(\mathbf{x}) >|^2}{||h||^2} \tag{5.2}$$

This maximization yields a feature, $\phi_{\hat{h}}(\cdot)$, that is optimally correlated to the perceptual results within the chosen class of features.

The class of signal features to use can be chosen based upon the specific application. The broader and more descriptive the class of features are, the better the fit will be to perception. Since the auditory system relies on time-frequency-like decomposition, and sonar echoes are time-varying in nature, we demonstrate this approach using features derived from a time-frequency representation (TFR).

5.2.1 Constrained Weighting Function

Expanding upon previous work by Atlas *et al.* [29, 30] as well as Davy *et al.* [31–33], we first demonstrate this approach using features identified from a generalized TFR. As Cohen showed [24], all quadratic TFRs can be represented by a two-dimensional Fourier transform of a weighted ambiguity function,

$$G_h(t,\omega) = \mathcal{F}^{-1}_{\theta \to t}\{\mathcal{F}_{\tau \to \omega}\{h(\theta,\tau)A(\theta,\tau)\}\} \tag{5.3}$$

where $G_h(t,\omega)$ is generalized quadratic TFR and $h(\theta,\tau)$ is delay-Doppler weighting function. By constraining h to be a function of only a few parameters, Davy *et al.* [31–33] were able to solve for the weighting function that produced a generalized TFR with maximal discrimination power.

Using this framework we define a class of features based on the energy contained in a generalized TFR,

$$\phi_h(x) = \sum_{t,\omega} G_h(t,\omega) \tag{5.4}$$

The extent of the weighting function in both delay and Doppler directly affects the energy content in the TFR. Therefore, our goal according to (5.2) is to find the weighting function, $\hat{h}(\theta,\tau)$, that produces a generalized TFR whose energy is maximally correlated to the MDS results.

Following Davy's approach we parameterize h to reduce the number of variables to be estimated. Restricting h to be a two-dimensional mixture of Gaussian distributions in delay-Doppler space,

$$h(\theta,\tau) = \sum_{i=1}^{N_{mix}} \exp\left[-\left(\left(\frac{\theta-\mu_{\theta i}}{\sigma_{\theta i}}\right)^2 - \rho_i\left(\frac{\theta-\mu_{\theta i}}{\sigma_{\theta i}}\right)\left(\frac{\tau-\mu_{\tau i}}{\sigma_{\tau i}}\right) + \left(\frac{\tau-\mu_{\tau i}}{\sigma_{\tau i}}\right)^2\right)\right] \tag{5.5}$$

we need only solve for the means and variances for each Gaussian term. With this set of parameters we maximize the correlation via constrained optimization.

5.2.2 Unconstrained Weighting Function

Another class of features derived from the time-frequency domain is a weighted spectrogram. This set of features is found based on a weighting function applied directly in the time-frequency domain. They are defined according to

$$\phi_h(x) = \sum_{t,\omega} h(t,\omega) S_x(t,\omega) \tag{5.6}$$

where $S_x(t,\omega)$ is the spectrogram of signal x and $h(t,\omega)$ is a time-frequency weighting function. With features of this form our approach will identify regions of time-frequency that are relevant to perception.

For this class of signal features, the optimal time-frequency weighting function can be solved for directly. Combining (5.6) with (5.2) we get

$$\hat{h}_k = \underset{h}{\operatorname{argmax}} \frac{\sum_i c_k(x_i) \cdot \sum_{t,\omega} h(t,\omega) S_{x_i}(t,\omega)}{\sum_{t,\omega} h^2(t,\omega)} \tag{5.7}$$

Since the denominator depends only on the energy in $h(t,\omega)$, the maximum can be obtained my maximizing the numerator subject to the assumption that the denominator is fixed and non-zero. Rearranging the summations and applying the Cauchy-Schwarz inequality to the numerator we arrive at

$$\begin{aligned} &\left| \sum_{t,\omega} h(t,\omega) \cdot \sum_i c_k(x_i) S_{x_i}(t,\omega) \right|^2 \\ &\qquad \leq \sum_{t,\omega} \left|h(t,\omega)\right|^2 \cdot \sum_{t,\omega} \left| \sum_i c_k(x_i) S_{x_i}(t,\omega) \right|^2 \end{aligned} \tag{5.8}$$

Equality is met when $h(t,\omega)$ is equal to an integer multiple of $\sum_i c^k(x_i) S_{x_i}(t,\omega)$. Therefore, the solution to (5.7) with minimum energy $h(t,\omega)$ is

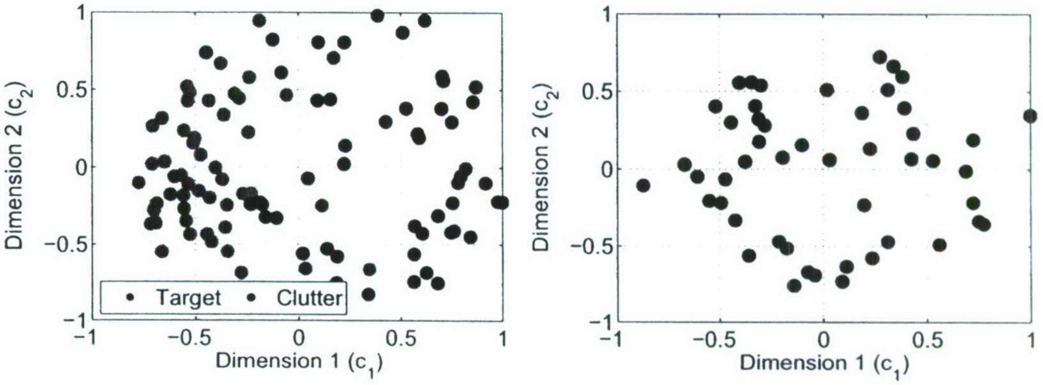
$$\hat{h}_k(t,\omega) = \sum_i c_k(x_i) S_{x_i}(t,\omega) \tag{5.9}$$

The optimally correlated weighting function is a perceptually-weighted sum of spectrograms.

This solution is advantageous as it provides a closed-form solution and therefore requires no constraining of the weighting function. While we have demonstrated this derivation using a time-frequency weighting function, the derivation also holds for a weighting function applied to any signal representation.

5.3 Example 1: U.S. Navy Data

Figure 5.1(a) shows a two-dimensional, nonmetric, MDS projection of the U.S. Navy signals using the perceptual similarity measures from Experiment 4 (Section 4.2.1). This figure shows strong separation between targets and clutter along the dimension 1 axis in this perceptual space. This visual insight confirms that if we do find physical signal attributes that correlate to this space, then those signals should produce reasonable classification features. Figure 5.1(b) shows a similar MDS space found using only the clutter similarity measures. The feature identification presented below will be based on the clutter-only MDS space due to restrictions on the target signals.



(a)

(b)

Figure 5.1: Two dimensional nonmetric MDS projection using a) the full U.S. Navy similarity matrix from Experiment 4 and b) the clutter-only similarity scores from the upper right quadrant of the similarity matrix.

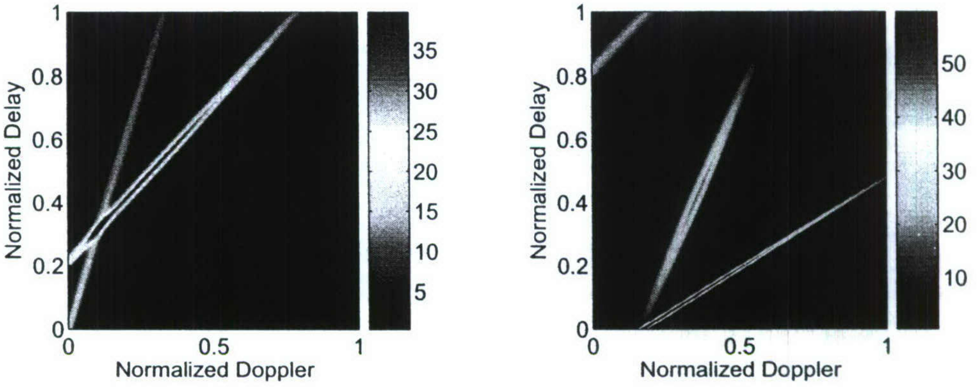
Standard Features

Using the traditional approach, we first attempt to identify if any standard features correlate well with each MDS dimension. We make use of a standard set of features that have proven to work well in past studies [1], [14], [49]. These features are defined in Appendix A. The correlation coefficients for these standard features against each MDS dimension is shown in Table A.1. The features exhibit a wide range of correlation values. The feature with the strongest correlation value is the maximum subband correlation frequency (maxSBCorrFreq) with a correlation value of 0.6068 to MDS dimension 1. This feature is a measure of the frequency location at which subband time envelopes are most correlated. Additionally, the feature with the highest correlation to dimension 2 is the frequency of minimum global subband attack time (MinGSATimeFreq), although this only correlated at a level of 0.3969.

This analysis suggests that subband correlation is a perceptually relevant feature. However, while being the best match to the MDS space, this feature only provides a correlation coefficient of 0.6068. This moderate level of correlation indicates that we have not found a highly meaningful representation of this perceptual space. The problem stems from the fact that this technique is highly dependent on the initial choice of features. Even with a set of 33 previously known features, only one correlated somewhat well to dimension 1 and none to dimension 2.

Generalized Time-Frequency

Optimal delay-Doppler weighting functions for MDS dimension 1 and 2 were found using a mixture of three Gaussian distributions. These weighting functions are shown in Figure 5.2. They isolate locations in ambiguity space that produce a generalized TFR whose energy is correlated to the MDS dimensions. Using these weighting functions we can identify features $\phi_{\hat{h}^1}(x)$ and $\phi_{\hat{h}^2}(x)$ via (5.4). Features found from these weighting function have correlation values of 0.6408 and 0.4682 to MDS dimensions 1 and 2, respectively. These correlations are not only higher than any individual feature, but required no *a priori* knowledge of potential feature sets.



(a)

(b)

Figure 5.2: Delay-Doppler mask for a) dimension 1 and b) dimension 2 of the U.S. Navy MDS analysis.

Time-Frequency Weighting Function

Figure 5.3 illustrates the optimal time-frequency weighting functions for MDS dimensions 1 and 2 found according to (5.9). These functions emphasize regions of time-frequency that play an important role in each MDS dimension. Specifically, sonar echoes that exist in negative regions of the time-frequency weighting function will tend towards the negative end of the MDS space and sonar echoes that exist in positive regions will tend toward the positive region of the MDS space. From the weighting functions, we see that dimension 1 appears to differentiate between relatively short duration high frequency signals and diffuse low-frequency signals. Similarly, dimension 2 reverses the roles of high and low frequency, differentiating between diffuse high frequency and short duration low frequency signals.

Using these weighting functions, we also identify signal features $\phi_{\hat{h}^1}(x)$ and $\phi_{\hat{h}^2}(x)$ via (5.6). These new features have a correlation coefficient of 0.8040 and 0.7646 to dimensions 1 and 2, respectively. These values are much higher than any single feature found from the standard feature set, indicating that these weighting functions not only give visual insight into the nature of the perceptual space but also identify good perceptual features.

5.4 Example 2: Boundary 2004 Data

A two-dimensional, nonmetric, MDS projection of the Boundary 2004 signals using the perceptual similarity measures from Experiment 5 (Section 4.2.2) is shown in Figure 5.4. This space reveals distinct perceptual clustering within each class. For example, one group of targets lies toward the far left of the plot, while another group lies towards the upper right. This perceptual distinction is most likely a result of the listeners distinguishing between different types of targets. In contrast to the U.S. Navy data, these signals require both dimension to best distinguish between classes.

Standard Features

Values for the standard feature set correlated against the Boundary 2004 MDS dimensions is shown in Table A.2. The feature with the strongest correlation, spectral rolloff (specRolloff), is a measure of how concentrated a signal's energy is towards low frequency. That is, it distinguishes between low and high

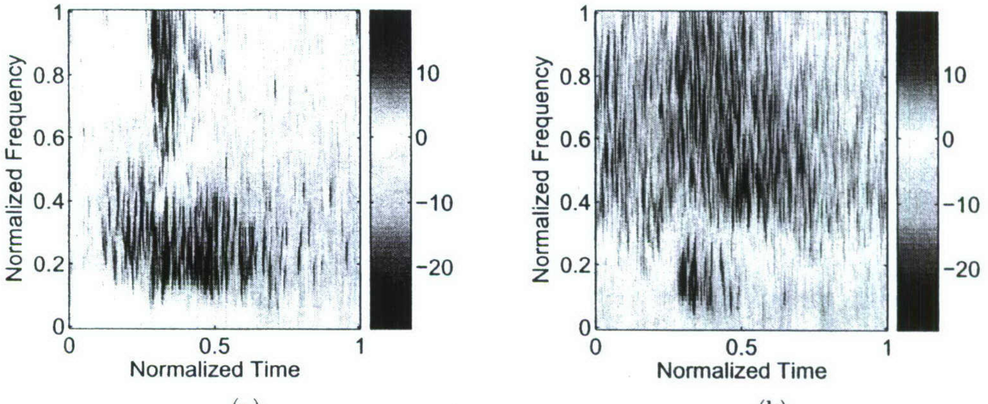


Figure 5.3: Perceptually-weighted spectrogram for a) dimension 1 and b) dimension 2 of the U.S. Navy MDS analysis (clutter-only).

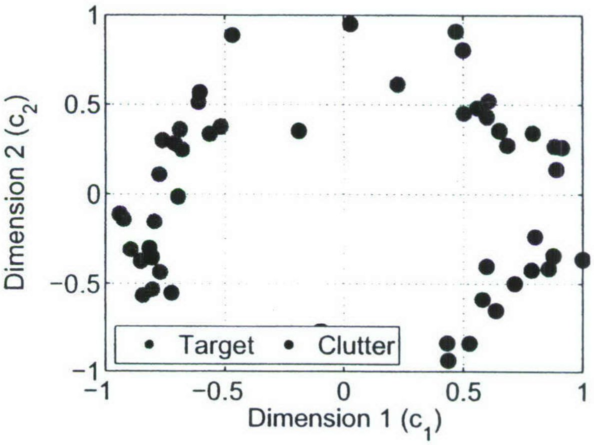


Figure 5.4: Two dimensional nonmetric MDS projection using the full Boundary 2004 similarity matrix from Experiment 5.

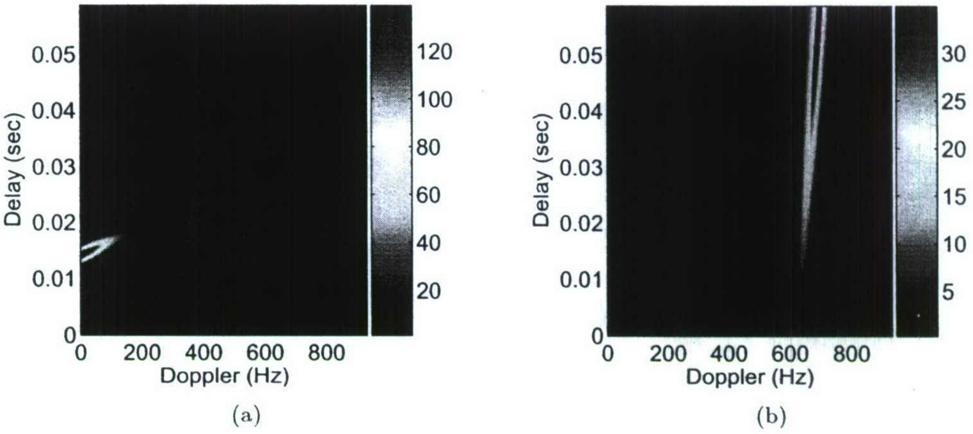


Figure 5.5: Delay-Doppler mask for a) dimension 1 and b) dimension 2 of the Boundary 2004 MDS analysis.

frequency signals. Spectral rolloff correlates with MDS dimension 1 at a level of -0.8205. The feature with the greatest correlation to dimension 2 is rise time with a value of 0.5449.

Generalized Time-Frequency

Optimal delay-Doppler weighting functions for MDS dimension 1 and 2 were again found using a mixture of three Gaussian distributions. They are shown in Figure 5.5. From these weighting functions, we derive features that correlate with MDS dimension 1 and 2 at a value of 0.7240 and 0.4108, respectively. In this case, these features came close, but did not exceed the correlation value for the best standard features. One reason for this may be that the delay-Doppler plane may be a improper space to work in for these types of sonar echoes. Another possible reason could be that the gaussian mixture model we imposed could be a incorrect constraint for the delay-Doppler weighting function.

Time-Frequency Weighting Function

Figure 5.6 shows the optimal time-frequency weighting functions for MDS dimension 1 and 2 using the Boundary 2004 data. From Figure 5.6(a) we infer that dimension 1 discriminates between low and high frequency signals. This frequency dependence is expected because the known feature Spectral Rolloff correlated well to this dimension. In addition to this frequency selectivity, the weighting functions also reveal that the group of target signals on the negative end of dimension 1 have a characteristic double echo as seen in the figure. The dimension 2 weighting function appears to isolate a very short duration high frequency echo from a longer lower frequency echo. Features derived from these functions correlate to MDS dimensions 1 and 2 at a level of 0.9160 and 0.7160, respectively. This correlation is higher then any previous feature.

5.5 Example 3: Modeled Data

The two-dimensional parameter space for the Modeled Mine data is shown in Figure 5.7(a). This space represents the physical composition of the underwater objects that were used to model these signals. We use this space in a similar manor to the MDS space, except in this case we are looking for signal features that describe the physical parameter space as opposed to the perceptual space.

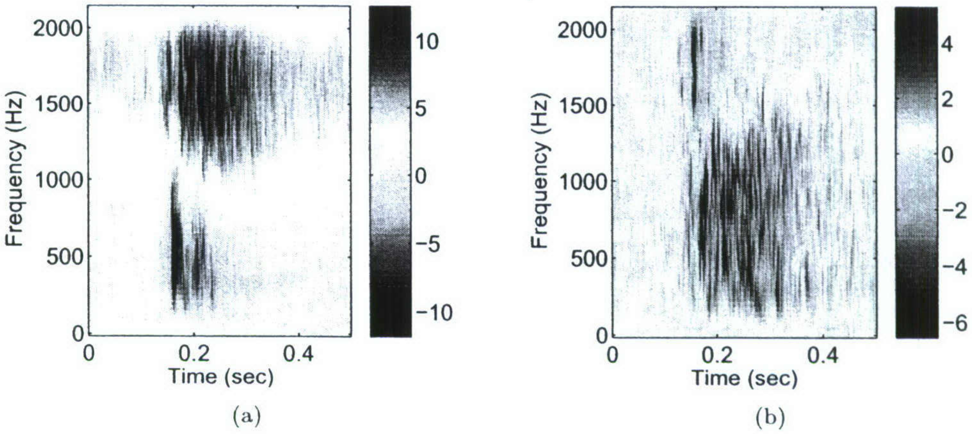


Figure 5.6: Perceptually-weighted spectrogram for a) dimension 1 and b) dimension 2 of the Boundary 2004 MDS analysis.

From this parameter space we notice that there is strong correlation between dimensions 1 and 2. To remove this correlation we employ Principle Component Analysis (PCA). PCA is used to rotate a space in order to isolate principle dimensions of variance. The rotated space after PCA is shown if Figure 5.7(b). Potential features will be correlated to these principle dimensions.

Standard Features

Values for the standard feature set correlated against the Modeled Mine parameter space are shown in Table A.3. Spectral Spread (specSpread) had the strongest correlation to dimension 1 with a value of 0.8521. The feature with the strongest correlation to dimension 2 is the frequency of minimum local subband attack slope (MinLSASlopeFreq). This correlated at a level of -0.3699.

Generalized Time-Frequency

Optimal delay-Doppler weighting functions for Modeled Mine parameter space were found using a mixture of three Gaussian distributions. They are shown in Figure 5.8. Features derived from these weighting functions correlated with each dimension at a value of 0.9065 and 0.4553.

Time-Frequency Weighting Function

Figure 5.9 shows the optimal time-frequency weighting functions for dimension 1 and 2 of the Modeled Mine parameter space. From these weighting functions we see a distinct resonance, or beat, pattern in time-frequency. The physical parameter space seems to distinguish between resonance structure located at low frequency and resonance structure located at high frequency. In addition to the time-frequency domain, the modulation domain (Section 2.2.1) is another space that efficiently isolates resonance structure. Therefore, we can demonstrate this approach on the modulation spectra of our signals as opposed to the spectrogram.

Using a modulation approach, our new class of features are defined as

$$\phi_h(x) = \sum_{\eta,\omega} h(\eta,\omega) M_x(\eta,\omega) \tag{5.10}$$

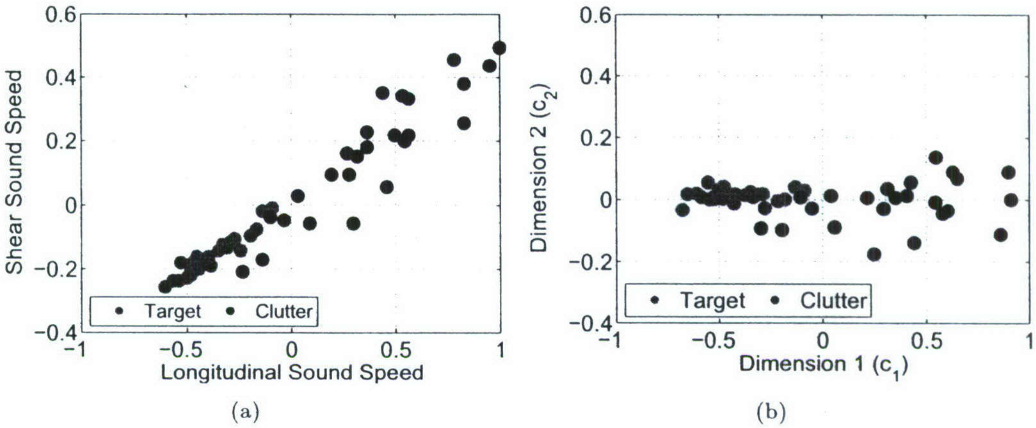


Figure 5.7: a) Parameter space of the Modeled Mine dataset and b) the rotated space after Principle Component Analysis.

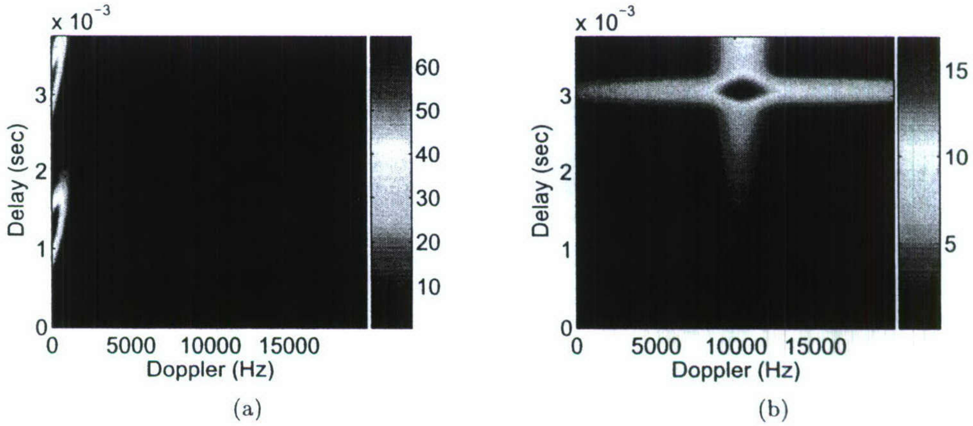
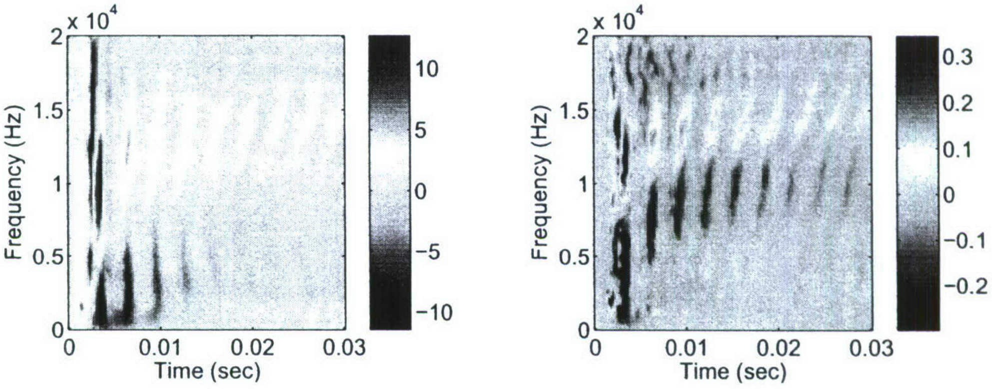


Figure 5.8: Delay-Doppler mask for a) dimension 1 and b) dimension 2 of the Modeled Mine parameter space (after PCA).

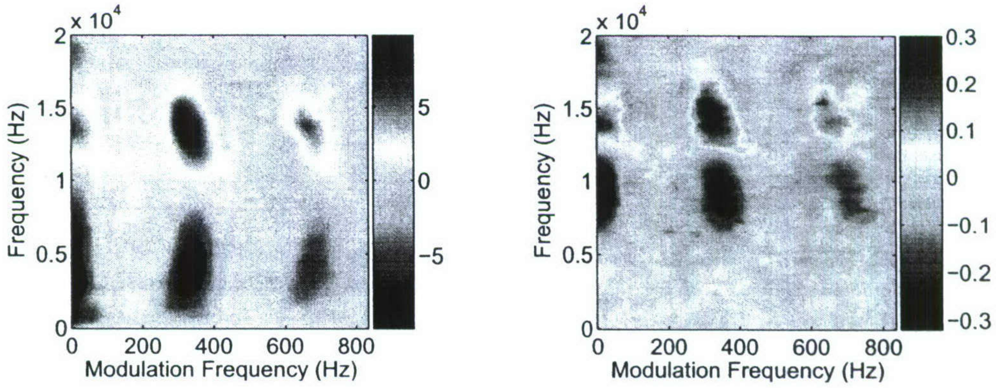
where $M_x(\eta, \omega)$ is the modulation spectra and $h(\eta, \omega)$ is a modulation-acoustic frequency weighting function. Following the same approach as before we can identify the optimal weighting function by a perceptually-weighted sum of modulation spectra.

Figure 5.10 shows the optimal modulation weighting function for each dimension of the parameter space. These figures again show a nice separation between low frequency resonance and high frequency resonance. The advantage of working in the modulation spectra is that this space is time-invariant. That is, it does not require time alignment of the signal to detect resonance patterns across the signals. Features derived from these modulation weighting functions correlate to dimension 1 and 2 of the parameter space at a level of 0.9187 and 0.3962, respectively.



(a) (b)

Figure 5.9: Perceptually-weighted spectrogram for a) dimension 1 and b) dimension 2 of the Modeled Mine parameter space (after PCA).



(a) (b)

Figure 5.10: Perceptually-weighted spectrogram for a) dimension 1 and b) dimension 2 of the Modeled Mine parameter space (after PCA).

Table 5.1: Summary of correlation values for each dataset

		Correlation to Dimension 1	**Correlation to Dimension 2**
U.S. Navy	Best Standard Feature	0.6068	0.3969
	Delay-Doppler Mask	0.6408	0.4682
	Time-Frequency Mask	0.8040	0.7646
Boundary 2004	Best Standard Feature	-0.8205	-0.5449
	Delay-Doppler Mask	0.7240	0.4108
	Time-Frequency Mask	0.9160	0.7160
Modeled Mine	Best Standard Feature	0.8521	-0.3699
	Delay-Doppler Mask	0.9065	0.4553
	Modulation-Frequency Mask	0.9187	0.3962

5.6 Discussion

We proposed a new approach for learning perceptually relevant features from listening experiment data. Traditionally, standard feature statistics are correlated against multidimensional scaling data to identify the feature with the greatest correlation. We introduced MDS matching in which an explicit maximization is performed to identify the best feature over a general class of signal features. This approach was demonstrated using two types of weighting functions: constrained and unconstrained.

The results from three different datasets are summarized in Table 5.1. The highlighted features represent the most correlated feature to each perceptual space. In each case the weighting functions outperform features drawn from a standard set. The MDS matching approach not only identified strong perceptual features without requiring *a priori* knowledge of candidate features, but also provided visual insight into the perceptual space.

Chapter 6

Kernel Identification

The methods outlined in the previous chapter first require multidimensional scaling before any perceptual features can be identified. This requirement adds an unnecessary layer of estimation. This chapter introduces an alternate approach of fitting directly to the similarity measures gathered from the listening experiments. We observe that the listener similarity matrix is a perceptual equivalent to matrices used in kernel methods for regression and classification such as kernel-based PCA [50] and Support Vector Machines [36]. In the following section we introduce kernel methods and how they relate to perceptual similarity. We then propose two methods for learning new kernel function from perceptual similarity data. Finally, we demonstrate this approach on the three sonar datasets.

6.1 Kernel Feature Space

Kernel methods use relational measures between data points for regression and classification instead of directly using signal features. These methods compute "similarity" between signals via a predefined kernel. These kernels are commonly defined as an inner product of the form

$$K(x_i, x_j) =< \phi(x_i), \phi(x_j) > {}^{1} \tag{6.1}$$

where $\phi(\cdot)$ is a mapping from input space to a feature space. The power of this approach is that the underlying feature space does not need to be defined explicitly, only the function that measures the relation between signals $K(x_i, x_j)$ is required.

Identifying an appropriate kernel function for a given problem is often a difficult task. There are many bivariate functions to choose from and there is no clear rule as to which will work best for a specific application. Three examples of commonly used kernels are polynomial kernels

$$K(x_i, x_j) = (< x_i, x_j > +c)^p , \tag{6.2}$$

hyperbolic tangent kernels

$$K(x_i, x_j) = \tanh(\kappa < x_i, x_j > +\Theta), \tag{6.3}$$

and most commonly the radial basis kernel

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} . \tag{6.4}$$

[1] It is important to note that this strict definition for a kernel, often known as a Mercer kernel, has been relaxed in modern kernel techniques to include any relational measure (*e.g.* Potential Support Vector Machines [40, 41]).

These kernels have all been shown to perform well in various studies, but the ultimate choice on which to apply remains trial and error.

Our perceptual model we used for feature identification fits direction into this framework of kernel based classification. We have assumed that a human listener extracts perceptually relevant acoustic cues (or features) from a signal whenever they are asked to make judgments in an experiment. While we do not have direct access to these features, we can record the similarity measures they provide in a listening experiment. By finding a functional form for these similarity judgements we can learn a perceptually appropriate kernel function without needing to know the exact acoustic cues that are being used. This kernel approach allows us to bypass estimating the perceptual space via MDS all together. In addition to this computational nicety, previous studies have indicated that humans may use relational measures rather than specific features when aurally classifying sounds [51].

6.2 Similarity Fitting

To learn a numeric function representing perceptual similarity, we follow our previous approach by regressing over a functional model for similarity. We employ the following model

$$\hat{h} = \underset{h}{\operatorname{argmin}} \left\| \delta(\mathbf{x}, \mathbf{x}') - d_h(\mathbf{x}, \mathbf{x}') \right\|^2 \tag{6.5}$$

where $\boldsymbol{\delta}$ is a perceptual similarity matrix and d_h is a numeric similarity matrix. With this approach, our goal is to learn a numeric similarity measure between signals that is as close as possible to the similarity measures provided by humans listeners. One way to view this is as a warping of a numeric feature space to a new space in which signals are arranged according to their perceptual similarity.

6.2.1 Linear Similarity Fitting

The goal of similarity fitting is to learn a distance metric in a numeric feature space $\boldsymbol{\phi}(\cdot)$ in which signals are separated according to perception. One linear model for this distance metric is

$$d_h(x_i, x_j) = \sum_{k=1}^{p} h_k \cdot |\phi_k(x_i) - \phi_k(x_j)|^2 \tag{6.6}$$

where h_k is a scaling factor for the k^{th} dimension of the numeric feature space and p is the total number of dimensions. This model is a weighted Euclidian distance measure in which each feature dimension can be weighted according to their fit to perceptual similarity.

While this weighted Euclidian distance allows each feature dimension to be scaled, it does not allow for cross-terms between dimensions. Another L_2 model for similarity with freedom to scale cross-terms is

$$d_H(x_i, x_j) = \boldsymbol{\phi}(x_i)^T \cdot H \cdot \boldsymbol{\phi}(x_j) \tag{6.7}$$

This similarity model is based on the Mahalanobis distance, where H is a $p \times p$ scaling matrix. This model for similarity is a weighted inner product between feature vectors. If H is set to the identity matrix then this measure is a standard inner product. In contrast to (6.6), this model assumes that perceptual similarity is numerically assigned such that larger numbers equate to greater similarity.

Using the Mahalanobis distance model, we now wish to learn the weights H that provides the best fit to perception, according to (6.5). In order to learn these weights, we first define a data matrix of training data as

$$\Phi = \begin{bmatrix} --- & \phi^T(x_1) & --- \\ --- & \phi^T(x_2) & --- \\ & \vdots & \\ --- & \phi^T(x_N) & --- \end{bmatrix}_{N \times p} \tag{6.8}$$

With this data matrix, we construct an $N \times N$ numeric distance matrix $d_H = \Phi \; H \; \Phi^T$. Substituting this distance matrix into (6.5) we get

$$\hat{H} = \underset{H}{\operatorname{argmin}} \left\| \boldsymbol{\delta} - \Phi \; H \; \Phi^T \right\|^2 \tag{6.9}$$

Using the Frobenius norm this minimization can be rewritten as

$$\hat{H} = \underset{H}{\operatorname{argmin}} \; \operatorname{tr} \left\{ \left(\boldsymbol{\delta} - \Phi \; H \; \Phi^T \right)^T \left(\boldsymbol{\delta} - \Phi \; H \; \Phi^T \right) \right\} \tag{6.10}$$

where tr is the trace of the matrix. We can now solving for the weights H by taking the derivative of (6.10) with respect to H

$$\frac{d}{dH} = 2 \; \Phi^T \; \Phi \; H \; \Phi^T \; \Phi \; - 2 \; \Phi^T \; \boldsymbol{\delta}^T \; \Phi \tag{6.11}$$

Setting (6.11) equal to zero and solving for H, we find

$$\hat{H} = \left(\Phi^T \; \Phi \right)^{-1} \Phi^T \; \boldsymbol{\delta} \; \Phi \left(\Phi^T \; \Phi \right)^{-1} \tag{6.12}$$

The above derivation provides a nice closed form solution for learning a perceptually-driven metric space. The derived distance measure $d_{\hat{H}}$ can now be used as a kernel function in kernel based signal classification and regression.

6.2.2 Nonlinear Similarity Fitting

A linear warping of signals in a feature space is not always enough to accurately describe recorded perceptual similarity. In this case a more complex nonlinear approach is required. As we do not have insight into all of the processes used by humans when judging perceptual similarity, it would be difficult to identify a specific nonlinear model to use. Therefore, we will take a nonparametric approach to nonlinear similarity fitting.

Assume we wish to measure the similarity between a new test signal x^* and training signal x_i. To measure this similarity, we first identify a set of training signals $\mathcal{G}$ that are the k-nearest-neighbors of x^*. We will estimate the similarity between x^* and x_i using a weighted sum of nearest-neighbors

$$\hat{\delta}(x^*, x_i) = \sum_{j \in \mathcal{G}} \beta_j \cdot \delta(x^*, x_j) \tag{6.13}$$

where β_j is a weighting coefficient for the j^{th} nearest-neighbor. The estimate of similarity between x^* and x_i is therefore a weighted average of similarities between x_i and the nearest-neighbors to x^*.

We choose the weighting coefficients such that the weights sum to one making (6.13) a true average. We also choose the weights such that the closer the nearest-neighbors the greater the weight. This is according to

$$\beta_j = \frac{\exp\{-\alpha \cdot d^2(x^*, x_j)\}}{\sum_{m \in \mathcal{G}} \exp\{-\alpha \cdot d^2(x^*, x_m)\}} \tag{6.14}$$

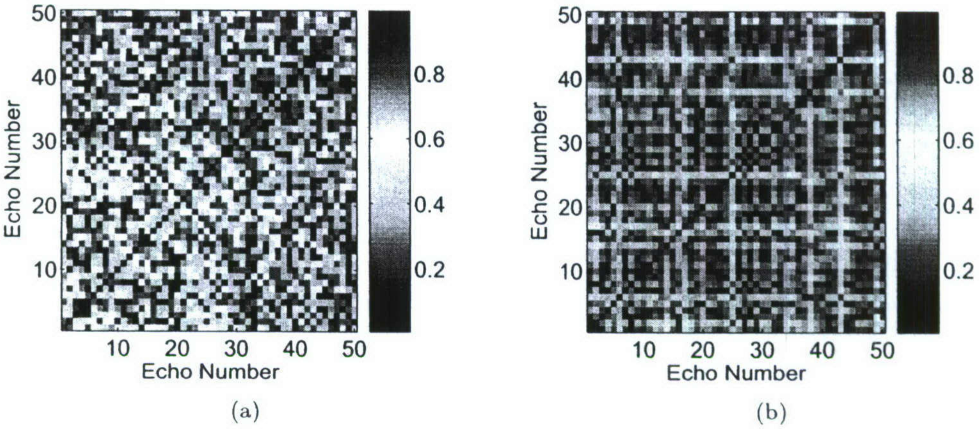


Figure 6.1: a) Clutter only similarity matrix from Experiment 4 and b) Euclidian distance matrix calculated from signal feature space.

where $d(\cdot,\cdot)$ is a numeric distance measure that is used to identify nearest-neighbors and α is a scaling factor.

The advantage of this technique is that it does not imposes any structure on the estimated similarity. It does not even require a numeric feature space $\boldsymbol{\phi}$. The technique only requires a distance measure between signals and it identifies which training signals are most like the test signal. It then leverages their known perceptual similarity ratings. The disadvantage of this approach is that it is dependent upon the training signal spanning the space of possible signals.

6.3 Example 1: U.S. Navy Data

In order to identify a functional distance metric that measures perceptual similarity we first calculate the standard Euclidian distance between signals. This distance measure not only quantifies a baseline to compare any further results, but it also provides a starting point for our regression techniques. To measure the distances between these signals, we first identify a simple yet descriptive feature set.

For the U.S. Navy dataset we choose to extract local time moments over each subband of a spectrogram. That is,

$$\langle t^n \rangle_\omega = \frac{\sum_t t^n \; S(t,\omega)}{\sum_t S(t,\omega)} \tag{6.15}$$

These moments describe the average time, duration, skew, *etc* of the echo at various frequency regions. We extract the first three moments, $n = \{1, 2, 3\}$, from a spectrogram with nine subbands. This resulting 27 dimensional feature space will be used as a point of comparison to the metric spaces found using similarity fitting.

Figure 6.1(a) shows the perceptual similarity matrix of the clutter-only signals from Experiment 4. As a comparison, Figure 6.1(b) show the Euclidian distance in feature space between the same set of signals. Both matrices have been normalized between zero and one, where zero is very similar and one is very different. The alignment, according to (2.22), between these matrices is 0.3398. Using similarity fitting we will attempt to learn a distance measure that is more accurately describes perceptual similarity.

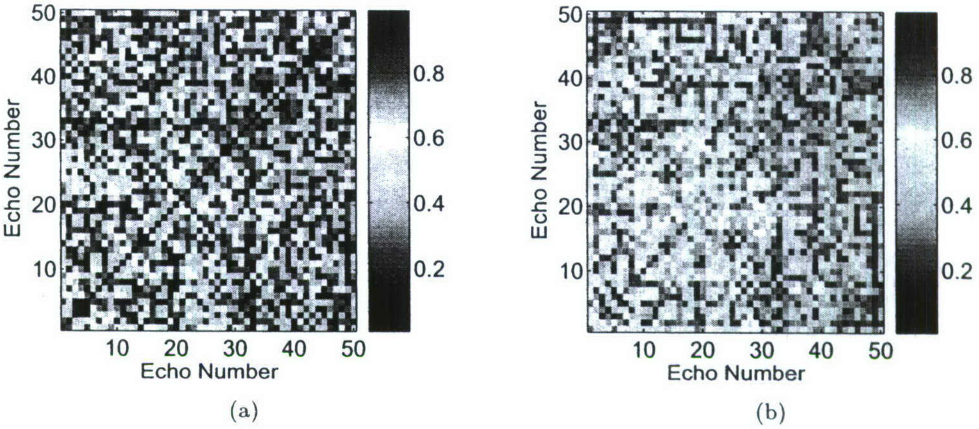


Figure 6.2: a) Clutter only similarity matrix from Experiment 4 and b) perceptual Mahalanobis distance matrix.

Mahalanobis Distance

To improve upon the Euclidian distance matrix, we seek to identify a perceptually appropriate Mahalanobis distance measure. Our Mahalanobis regression technique requires that our similarity matrix is scaled such that larger numbers equate to a greater degree of similarity. To do this, we simply take one minus the values of the similarity matrix shown in Figure 6.1(a). Next we construct a data matrix Φ using the local moments of the 50 U.S. Navy clutter echoes. With this data matrix, a scaling matrix $\hat{H}$ can be found via (6.12).

A perceptual Mahalanobis distance matrix of the U.S. Navy clutter signals is now constructed using the scaling matrix $\hat{H}$. Figure 6.2 shows a comparison of one minus the Mahalanobis distance matrix $(1 - d_{\hat{H}})$ against the perceptual similarity matrix. The Mahalanobis distance matrix is aligned with the similarity matrix at a level of 0.8360. This is much higher than the correlation with the Euclidian distance matrix.

Nearest-Neighbor

Another approach to improving upon the Euclidian distance metric is Nearest-Neighbor distance regression. To identify a similarity matrix for the clutter signals using this approach we employ a leave-one-out cross validation strategy [52]. First we designate a testing echo from the set of 50 clutter echoes. Nearest-neighbors are then calculated from the feature space of local moments. Next, the test echo's similarity to all other echoes is calculated using (6.13). This process is repeated for every clutter echo in the dataset.

Figure 6.3 shows a comparison of the nearest-neighbor similarity matrix with the perceptual similarity matrix. The alignment between these matrices is 0.5402. This is again higher than the Euclidian distance metric.

6.4 Example 2: Boundary 2004 Data

Following the same approach as with the U.S Navy data, we look to identify a perceptual distance metric that equates to the Boundary 2004 perceptual similarity matrix from experiment 5. We again calculate 27 local moments of the spectrogram to provide us with an initial feature space. Figure 6.4 shows the

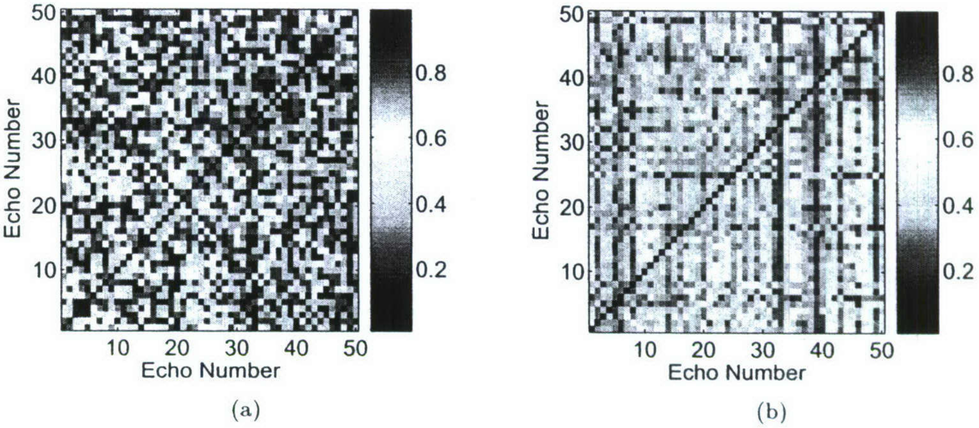


Figure 6.3: a) Clutter only similarity matrix from Experiment 4 and b) numeric similarity matrix after nearest-neighbor similarity fitting.

comparison of a Euclidian distance matrix found in this features space with the perceptual similarity matrix from Experiment 5. The alignment between these two matrices is 0.6079.

Mahalanobis Distance

A perceptual Mahalanobis distance for the Boundary 2004 signals is found in the same manner as with the U.S. Navy signals. Figure 6.5 shows a comparison of the Mahalanobis distance matrix against the perceptual similarity matrix. The Mahalanobis distance matrix aligns with the similarity matrix at a level of 0.9329. As with the U.S. Navy data, this correlation is much higher that a simple Euclidian distance metric.

Nearest-Neighbor

Following the same approach as the U.S. Navy data we identify a numeric perceptual similarity metric for the Boundary 2004 data using Nearest-Neighbor distance regression. The results of this regression are shown in Figure 6.6. The alignment between the perceptual matrix and the nearest-neighbor matrix is 0.8104.

6.5 Example 3: Modeled Data

In the case of the Modeled Mine Data, we attempt to learn a distance metric between echoes that equates to physical distance between the objects that the echoes were modeled from. That is, we wish to identify a distance metric for the distance between between objects in parameter space. As before we calculate an initial set of standard features, however in this case the features we choose are the first ten cepstral coefficients, c_{1-10} (2.5). These ten features provide an initial feature space to compare against our later results.

Figure 6.7(a) shows a Euclidian distance matrix between signals in the Modeled mine database calculated in parameter space. Signals 1-25 are target echoes and 26-50 are clutter echoes. In contrast, Figure 6.7(b) show a Euclidian distance matrix between the same signals but calculated in the cepstral feature space. This matrix looks to contain less separation between target and clutter echoes. The alignment

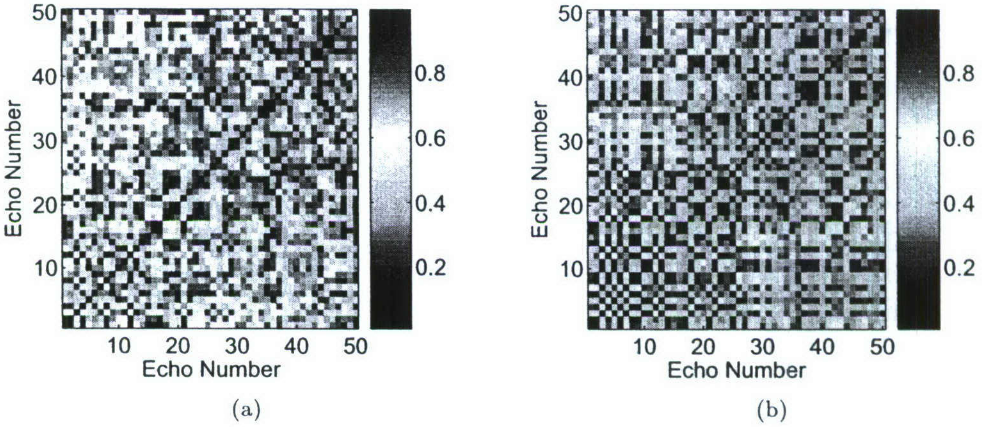


Figure 6.4: a) Similarity matrix from Experiment 5 and b) numeric similarity matrix calculated from Boundary 2004 signal feature space.

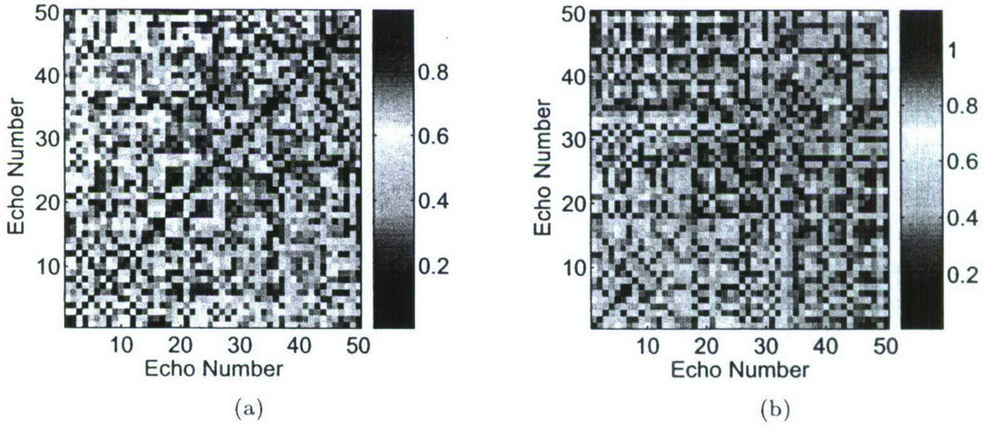


Figure 6.5: a) Similarity matrix from Experiment 5 and b) numeric similarity matrix after Mahalanobis similarity fitting.

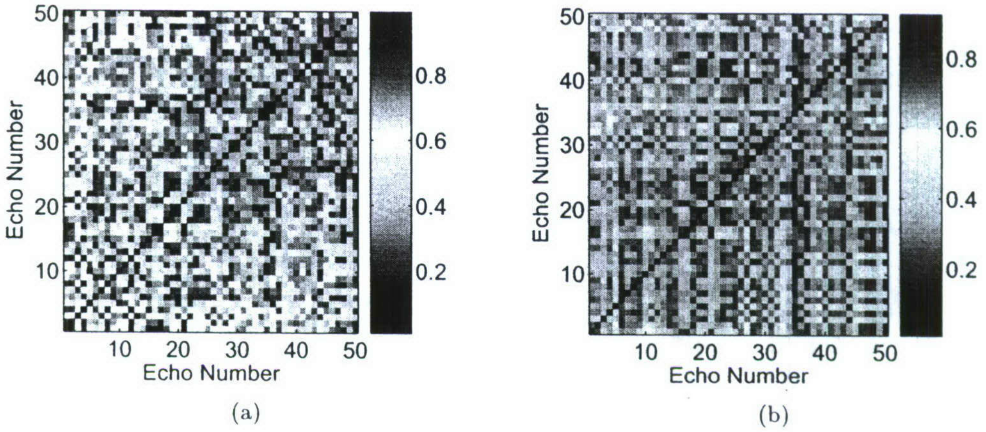


Figure 6.6: a) Similarity matrix from Experiment 5 and b) numeric similarity matrix after nearest-neighbor similarity fitting.

between these two matrices is 0.5349. Next we apply our two regression techniques to identify a more accurate distance metric for distance between objects in parameter space.

Mahalanobis Distance

A Mahalanobis distance for the Modeled Mine data is found in the same manner as with the U.S. Navy and Boundary 2004 signals. As opposed to the previous examples, the Mahalanobis distance describes physical distance between objects in parameter space rather than perceptual distance. Figure 6.8 shows a comparison of the Mahalanobis distance matrix against the parameter distance matrix. The Mahalanobis distance matrix aligns with the similarity matrix at a level of 0.7858.

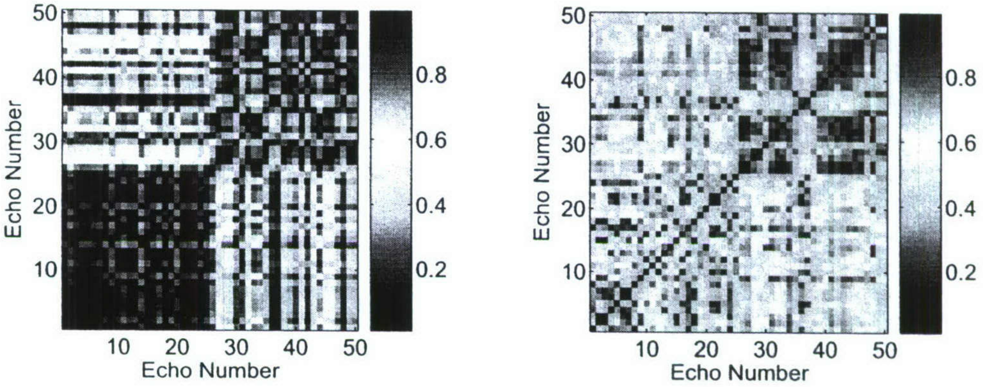
Nearest-Neighbor

Using Nearest-neighbor distance regression we now find a non-linear distance measure that fits to the physical parameter distance. Following the same approach as before, we identify a nearest-neighbor distance matrix for the signals in the modeled mine data. Figure 6.9 shows this matrix compared to the parameter distance matrix. The alignment between these two matrices is 0.9659. This correlation indicates that this regression provides a very good fit to distance between objects in physical parameter space.

Classification

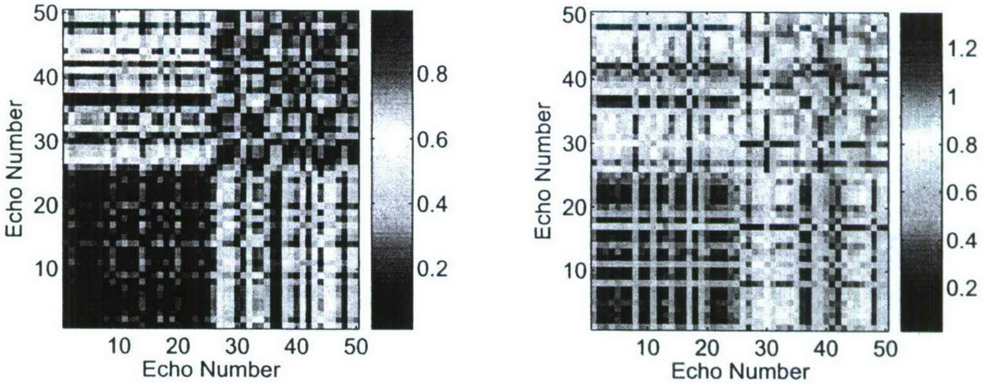
We now compare the classification ability of the three distance metrics for the Modeled Mine data (Euclidian, Mahalanobis, Nearest-Neighbor) using the kernel-based potential support vector machine (pSVM) classifier [40, 41]. The potential support vector machine is a generalization of the more common support vector machine. The advantage of the pSVM is that it does not require a Mercer kernel – it allows any relational measure to be used as the kernel function.

To test the distance metrics found in the above sections we model a new set of testing data. This data is drawn from the same distribution as the training data (Section 3.2.3) and consists of 100 target echoes and 100 clutter echoes. We train three different pSVM models using the above training data and each of the distance metrics. A class score is determined for each signal in the testing dataset using each pSVM



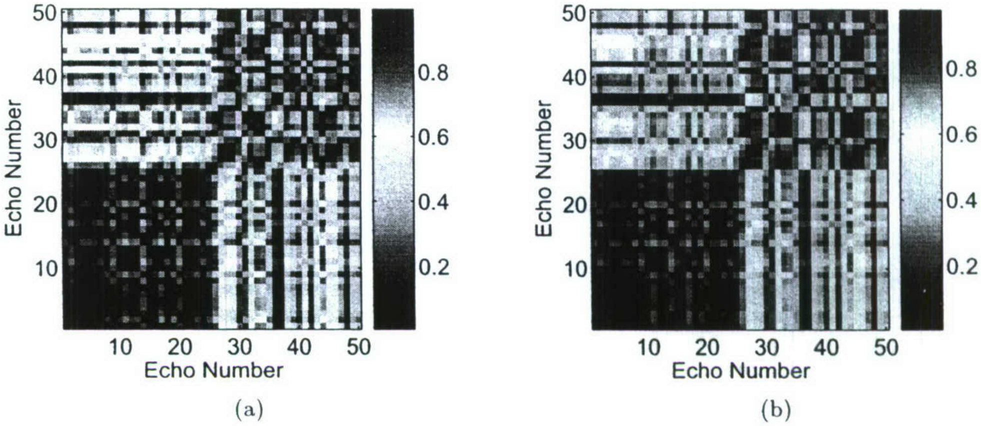
(a) (b)

Figure 6.7: a) Similarity matrix calculated from the Modeled Mine parameter space and b) numeric similarity matrix calculated from Modeled Mine signal feature space.



(a) (b)

Figure 6.8: a) Similarity matrix calculated from the Modeled Mine parameter space and b) numeric similarity matrix after Mahalanobis similarity fitting.



(a)

(b)

Figure 6.9: a) Similarity matrix calculated from the Modeled Mine parameter space and b) numeric similarity matrix after nearest-neighbor similarity fitting.

Table 6.1: Summary of correlation values of numeric similarity matrices to the perceptual similarity matrices

	Euclidian Distance	**Mahalanobis Distance**	***k*-Nearest-Neighbor**
U.S. Navy	0.3398	0.8360	0.5402
Boundary 2004	0.6079	0.9329	0.8104
Modeled Mine	0.5349	0.7858	0.9659

model. From this class score we can create ROC curves to compare the performance of each distance metric.

The classification results for each of the pSVM models are shown in Figure 6.10. The distance metrics that were found to fit to the parameter space perform notably better that the Euclidian distance measure found from a cepstral feature space. This improvement can be seen by the increased probability of correct detection for a given probability of false alarm. The nonlinear nearest-neighbor approach which provided the best correlation to the physical distance metric also resulted in the best classification ability. This demonstrates how a substandard feature space (cepstral) can be improved upon by fitting a physical (perceptual) distance metric.

6.6 Discussion

Kernel methods for regression and classification describe signals not based on a set of features, but on relational measures between signals. In this section, we demonstrated how this description of signals is analogous to the results of perceptual similarity experiments that are commonly used to identify signal features. We then proposed two methods for learning a perceptual kernel (distance metric) that depict how humans relate signals to one another. Mahalanobis distance regression identified this kernel using a linear parametric regression while Nearest-Neighbor distance regression identifies the kernel using a completely nonlinear nonparametric approach.

The results from these regression techniques on each dataset is summarized in Table 6.1. The distance metrics with the highest correlation values are highlighted. In each dataset, improvement was made over a standard Euclidian distance metric. In the Modeled Mine dataset we also showed that this improvement in distance metric also led to an improvement in classification ability.

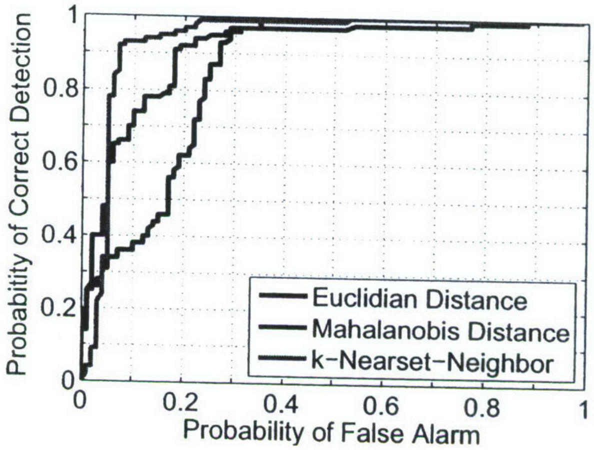


Figure 6.10: Modeled Mine ROC curves for different kernel (distance) matrices using the potential support vector machine

Chapter 7

Perceptual Prior

The listening experiments conducted in Chapter 4 showed that subjects can distinguish between sonar target and clutter echoes. In post-experiment discussions, many subjects felt that certain echoes did not seem to belong to their prescribed class. That is, they reported some targets sounding more like clutter and some clutter sounding more like a target. This disagreement between the subjects and class labels could be due to the fact that those particular sounds were unusual or outlier echoes or that they were simply mislabeled when the data was collected. These echoes have the potential to affect the training of an automatic classifier. In this chapter we discuss a method for using this obtained perceptual information to aid in the training of a classifier model. This new method uses a perceptual prior, which is defined as the probability a human listener would associate a signal with a particular class. We introduce the perceptual prior in the area of Bayesian decision theory and show how it can be used to aid in the estimation of the likelihood function for each class.

7.1 Bayesian Decision Theory

Bayesian decision theory is a fundamental statistical approach to the problem of automatic classification [53]. This approach relies on identifying probability distributions for each class. That is, we seek to identify the probability of class y_j given a set of signal features, $p(y_j|\phi(x))$. This probability is referred to as the posterior probability because it measures the probability of a given class after we have seen the data. A decision is made by simply choosing the class which has the greatest posterior probability

$$\hat{y}_j = \underset{y_j}{\operatorname{argmax}}\, p(y_j|\phi(x)) \tag{7.1}$$

In order to identify the posterior probability, Bayesian decision theory is based on Bayes' formula, which states

$$p(y_j|\phi(x)) = \frac{p(\phi(x)|y_j) * p(y_j)}{p(\phi(x))} \tag{7.2}$$

This formula shows that by observing the value of $\phi(x)$ we can convert the prior probability $p(y_j)$ to the posteriori probability $p(y_j|\phi(x))$. Referred to as the likelihood function, the probability $p(\phi(x)|y_j)$ is the key important term that connects the class probability before we have seen a signal (prior) to the class probability given a specific signal. The probability of the features $p(x)$, known as the evidence factor, can be viewed as merely a scale factor that ensures that the posteriori probabilities sum to one over all classes.

7.2 Likelihood Estimation

An accurate estimation of the likelihood probability density function (PDF) is essential in order to make an accurate class decision based on the posterior. To estimate this PDF, a set of training data is gathered which includes both signals and their associated class labels. From this data, a feature space $\phi(x)$ is calculated to reduce the dimensionality (and complexity) of the PDF that is being estimated. Next, the PDF is estimated from the data or its statistics. These statistics are calculated for each class and for each dimension. If we were to assume a normal distribution for that data, then the likelihood function can be estimated as

$$p(\phi_k(x)|y_j) \sim \mathcal{N}\left(\mu_{j,k}, \sigma_{j,k}\right) \tag{7.3}$$

In order to estimate the mean and variance statistics of a set of training data the most common estimators are

$$\mu_{j,k} = \frac{1}{N_j}\sum_i \phi_k(x_i) \cdot I_i^{(y_j)} \tag{7.4}$$

$$\sigma_{j,k}^2 = \frac{1}{N_j - 1}\sum_i \left(\phi_k(x_i) - \mu_{j,k}\right)^2 \cdot I_i^{(y_j)} \tag{7.5}$$

where $I_i^{(y_j)}$ is an indicator function for signals of class y_j and N_j is the number of signals in class y_j. Though these estimators assume that the training data from each class was uniformly sampled, this assumption is not always valid. For example, the training data is not always collected in the same environment in which the end classifier will operate in. This variation in data collection could lead to outlier signals in the training data. Another imbalance could be due to inconsistencies with the class labeling of the training data. That is, a true clutter echo may be labeled as a target or vice-versa. This non-uniformity of the training data was often reported by our human listeners. They felt that some echoes did not seem to belong to their prescribed class.

We introduce the perceptual prior to account for inconsistencies in the training data. The perceptual prior is the probability of a signal being a member of a particular class as determined by human subjects. This probability will be denoted as $p_j(x) = p(x \in y_j)$. Using the perceptual prior we define new mean and variance estimators as

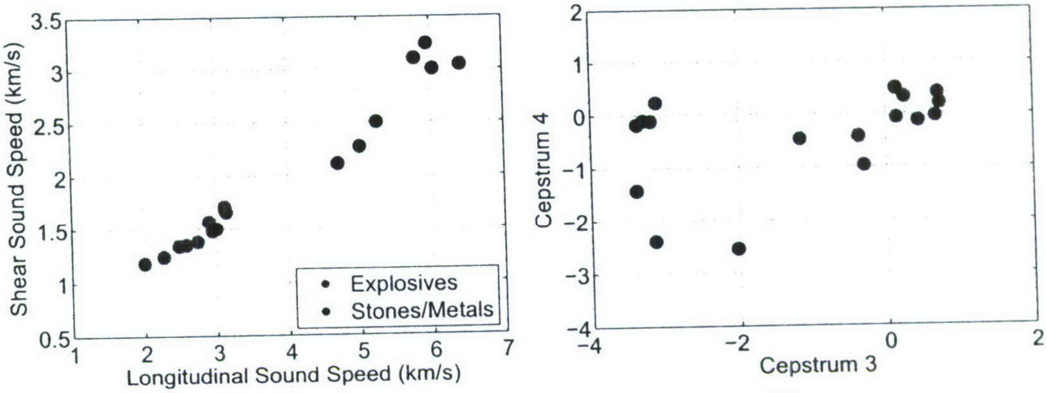
$$\mu_{j,k} = \frac{1}{\sum_l p_j(x_l)}\sum_i \phi_k(x_i) \cdot p_j(x_i) \tag{7.6}$$

$$\sigma_{j,k}^2 = \frac{1}{\sum_l p_j(x_l)}\sum_i \left(\phi_k(x_i) - \mu_{j,k}\right)^2 \cdot p_j(x_i) \tag{7.7}$$

These new estimators now weight each training signal based on how confident human listeners are that the signal arose from the class of interest. We illustrate this approach using modeled mine data in the following section. We illustrate how the perceptual prior can be identified and how it affects the target detection and false alarm rates on a set of test data.

7.3 Example: Modeled Data

We demonstrate how the perceptual prior affects classification using the modeled mine data (Section 3.2.3). In the modeled data, the perceptual prior is not identified from human listening experiments, but instead



(a)

(b)

Figure 7.1: The location of a set of training signals in a) parameter space and b) feature space.

is identified from the parameter space used to model the data. The parameter space is analogous to the perceptual MDS space that was identified from our listening experiment in Section 5. Both spaces provide insight into the true underlying arrangement of the signals that may not be evident in a suboptimal feature space.

Figure 7.1(a) shows the location in parameter space of 18 training signals. These training signals represent the acoustic properties of various kinds of explosive materials as well as stones and metals. Note that one stone echo seems to be located near the other explosives. This stone is limestone, which has acoustic properties that are very similar to many kinds of explosives. A sonar echo received from a limestone-filled sphere would be very difficult to distinguish from an explosive-filled sphere. Figure 7.1(b) shows the location for these same 18 signals in a two-dimensional feature space. The features represented are the second and third cepstral coefficients. These features define the space that can be measured when a new test signal is received, as such, this space is where the decision boundary must be drawn.

In order to identify the perceptual prior for training signal x_i we first estimate a PDF in parameter space for each class. This estimation is done excluding the signal of interest x_i. For the modeled mine case we simply estimate a Gaussian distribution for each class without using signal x_i in that estimation. Once these distributions are found, the perceptual prior is defined according to

$$p_j(x_i) = \frac{f_{y_j}(x_i)}{\sum_j f_{y_j}(x_i)} \tag{7.8}$$

where $f_{y_j}(\cdot)$ is the PDF for class y_j as estimated in the "perceptual" parameter space. This process is then repeated for each signal in the training dataset.

Using the feature space for the training data, we now estimate the likelihood function for the explosive class and the stone/metal class. Figure 7.2(a) overlays two estimated Gaussian distributions on top of the data, one for each class. These Gaussians were estimated according to (7.4) and (7.5). This figure shows significant overlap between the two likelihood functions, mostly due to limestone being included in the training set. Alternately, we can estimate these Gaussian distributions according to (7.6) and (7.7) using the perceptual prior. These Gaussian distributions are shown in Figure 7.2(b). This figure shows a much stronger separation between the two likelihood functions due to the fact that limestone's perceptual prior for the stone/metal class was low.

Figure 7.3 shows both the parameter space and the feature space for a set of testing data. We identify

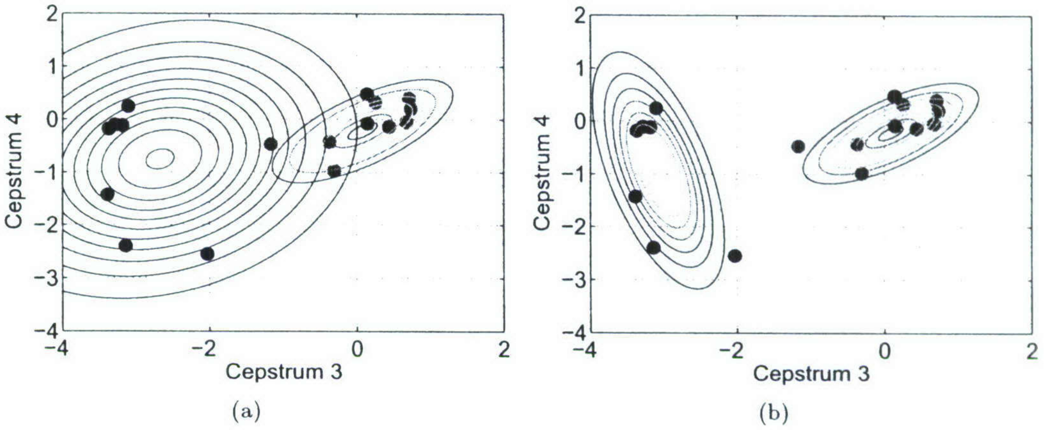


Figure 7.2: Estimated Gaussians for the likelihood distributions of explosives and stones/metals using a) standard estimators and b) perceptually weighted estimators.

Bayes decision boundaries for this test data using the likelihood functions estimated from the training data and assuming the priors are uniformly distributed (*i.e.* $p(y_1) = p(y_2) = .5$). Figure 7.4 shows two decision boundaries, one found without the perceptual prior and one found using the perceptual prior. The change in the likelihood estimate for the stone/metal class dramatically changes the boundary between the two classes. The perceptually-weighted model appears to capture the difference between the two classes. Without the perceptual model the Bayes decision boundary has a classification error rate of 7%; using the perceptual prior the error rate decreased to 2%.

7.4 Discussion

Bayesian decision theory fundamentally relies on an accurate estimate of the likelihood function for each class. Outlier and/or mislabeled signals that exist in the training dataset can often dramatically affect the likelihood estimate. In this chapter we proposed using human listener information in the form of a perceptual prior in order to better estimate the likelihood function. We demonstrated this approach on a set of modeled data, which through its parameter space, provided similar information as would a listening experiment. With a better estimate of the likelihood function we were able to decrease the classification error rate from 7% to 2%.

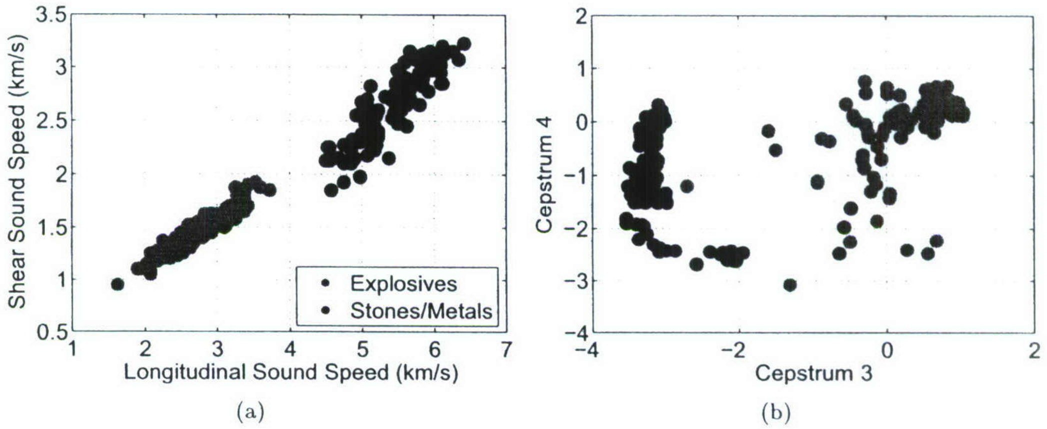


Figure 7.3: The location of a set of testing signals in a) parameter space and b) feature space.

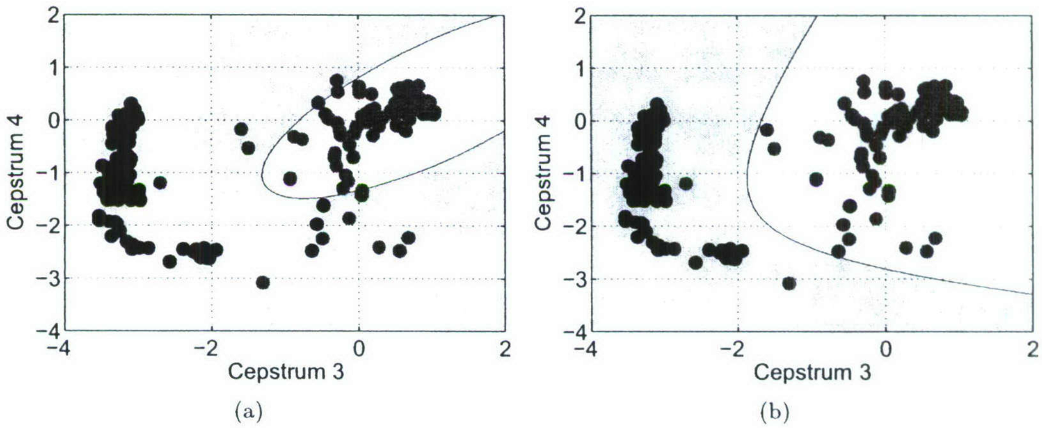


Figure 7.4: Decision boundaries overlayed onto the testing data for a) standard model b) perceptually-weighted model.

Chapter 8

Conclusions

The research discussed in this paper has developed a framework for employing perceptual information from human listening experiments to improve automatic event classification. We focus on the identification of new signal attributes, or features, that are able to predict the human performance observed in formal listening experiments. Using this framework, our newly identified features have the ability to elevate automatic classification performance closer to the level of human listeners. Below is an outline of the major contributions that were covered in this paper as well as a list of possible future directions for this work.

8.1 Contributions

The practical contributions of this paper are:

- **Listening Experiments** - Previously there had been anecdotal evidence that human sonar operators can distinguish active sonar target echoes from clutter echoes using aural cues alone. We conducted a series of listening experiments that validated this conventional wisdom. We also demonstrated that naïve listeners can be trained to perform this task comparable to trained navy sonar operators.

- **Feature Identification** - Psychoacoustics has always relied on simple hypothesis testing for the identification of perceptually relevant features. Previously, a known feature was compared against the results of a listening experiment to determine if that feature is relevant to perception. We presented a new method for learning a feature transform that maximizes the correlation between a feature and listening experiment results. These perceptually-driven feature transforms have the ability to provide a visual insight into how humans perceive sounds in a dataset.

- **Kernel Identification** - We presented a new approach for learning a kernel function by using the results of a similarity listening experiment. We introduce two new similarity-based regression techniques, one based on the Mahalanobis distance and one based on a local linear regression. These methods transform a feature space such that signals are arranged according to perceptual distances. These methods were shown to improve classification using a potential support vector machine.

- **Perceptual Prior** - Bayesian decision theory relies on accurately estimating the likelihood function for each class. We presented a method for improving the estimation of the likelihood functions when there are outliers and/or mislabeled data in the training dataset. This method involves using a perceptual prior as a measure of confidence that a signal is a member of a particular class. This perceptual prior was shown to provide a better decision boundary for the modeled mine data and thus increased classification performance.

8.2 Future Work

A number of methods have been proposed in this paper for utilizing perceptual information in the automatic classification of acoustic events. Future directions for this work are summarized as follows:

- In Chapter 4 we have shown that human sonar operators can distinguish between active sonar target and clutter echoes using impulsive charges as the source signal. These types of sources have a very wide bandwidth and short duration. Another type of source is a coherent signal (*e.g.* a chirp function). There have not been any studies to determine whether humans can aurally classify echoes using coherent sources and if so, what the frequency requirements (*i.e.* center frequency and bandwidth) of those sources are. There may be also be preprocessing requirements such as match filtering, frequency shifting or time scaling that is needed for auralization.

- We have identified a number of new signal features using MDS matching in Chapter 5. These features have been shown to fit the results of a listening experiment better than standard features. A problem with these time-frequency weighting functions is that they require accurate time alignment between the acoustic events. Therefore these weighting functions may be operationally impractical. Therefore we can simply use the time-frequency weighting functions as a guide to feature identification. Visually, the masks allow the researcher to determine what frequency range and time extent is used in perception. With this insight, more robust time and/or frequency features could be designed. These features could then be to evaluated in a cross-validated automatic classification experiment.

- In Chapter 5 we introduced MDS matching where perceptually-driven signal features were identified by preforming a linear regression between spectrograms and the MDS dimensions. This regression allowed us to identify regions in time-frequency that were relevant to perception. This approach could be extended to a nonlinear regression in order to model more complicated relationships. Features identified by a nonlinear regression have to possibility to provide a better fit between signals and perception. One possible model for this approach could be a generalized linear model where a nonlinear warping of the MDS space is performed prior to matching.

- When an automatic classifier is used in a new environment, signal characteristics can often change dramatically. Signals of the same class may no longer be close in feature space. In contrast, humans are known to be robust to changes in the channel response. Signals from the same class are perceived as similar regardless of the environment they were recorded in. If the perceptual kernel functions that were found in Chapter 6 truly reflect perceptual similarity, then they have the potential to be environmentally robust. This ability of the perceptual kernel will need to be evaluated on a dataset of sonar echoes from varying environments.

- In Chapter 7 we identified a perceptual prior that is used to estimate the likelihood function. This perceptual prior was found by estimating distributions in the parameter space from the modeled data. This rule is not necessarily optimal. A more direct strategy would be to conduct a listening experiment in which subjects were directly asked to assign a class membership probability to each signal. This strategy could provide a more accurate perceptual prior and therefore a better estimate of the likelihood function.

Bibliography

[1] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, vol. 61, 1977.

[2] W. A. Yost, *Fundamentals of Hearing*. Academic Press, 2000.

[3] S. McAdams and A. Bregman, "Hearing musical streams," *Computer Music Journal*, vol. 3, no. 4, pp. 26–43, 1979.

[4] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. Robert E. Krieger Publishing Co., 1974, (reprint of 1966 ed).

[5] J. A. Swets, "Is there a sensory threshold?" *Science*, vol. 134, pp. 168–177, 1961.

[6] J. A. Swets, W. Tanner, and T. Birdsall, "Decision processes in perception," *Psychological Review*, vol. 68, pp. 301–340, 1961.

[7] L. O. Harvey, "The critical operating characteristic and the evaluation of the expert judgment," *Organizational Behavior and Human Decision Processes*, vol. 53, no. 2, pp. 229–251, 1992.

[8] "http://www.rad.jhmi.edu/jeng/javarad/roc/jrocfiti.html." [Online]. Available: http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html

[9] H. Terasawa, M. Slaney, and J. Berger, "Perceptual distance in timbre space," in *International Conference on Auditory Display*, 2005.

[10] J. M. Grey, "Perceptual effects of spectral modifications on musical timbres," *Journal of the Acoustical Society of America*, vol. 63, 1978.

[11] J. H. Howard, "Psychophysical structure of eight complex underwater sounds," *Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 149–156, 1977.

[12] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbre," *Journal of the Acoustical Society of America*, vol. 94, 1993.

[13] J. Marozeau, A. de Cheveigne, S. McAdams, and S. Winsberg, "The dependency of timbre on fundamental frequency," *Journal of the Acoustical Society of America*, vol. 114, pp. 2946–2957, 2003.

[14] S. McAdams, "Perspectives on the contribution of timbre to musical structure," *Computer Music Journal*, vol. 23, 1999.

[15] S. Tucker and G. J. Brown, "Classification of transient sonar sounds using perceptually motivated features," *IEEE Journal of Oceanic Engineering*, vol. 30, pp. 588–600, 2005.

[16] M. L. Davidson, *Multidimensional Scaling.* Krieger, 1992.

[17] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications.* Springer, 1997.

[18] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," IRCAM, Tech. Rep., 2004.

[19] P. Loughlin and L. Cohen, "Moment features invarient to dispersion," *Proceedings of SPIE*, vol. 5426, pp. 234–246, 2004.

[20] J. Dugundji, "Envelopes and pre-envelopes of real waveforms," *IEEE Transactions on Information Theory*, vol. 4, no. 1, pp. 53– 57, 1958.

[21] J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification," *Journal of the Audio Engineering Society*, vol. 52, pp. 724–739, 2004.

[22] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, J. Bonnell, Ed. Prentice Hall PTR, 2001.

[23] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications.* Cambridge University Press, 1998.

[24] L. Cohen, *Time-Frequency Analysis*, A. V. Oppenheim, Ed. Prentice-Hall PTR, 1995.

[25] M. Vinton and L. Atlas, "A scalable and progressive audio codec," in *IEEE ICASSP*, 2001.

[26] R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions of Information Theory*, vol. 38, pp. 713–719, 1992.

[27] N. Saito and R. R. Coifman, "Local discriminant bases," in *Wavelet Applications in Signal and Image Processing II, Proc. SPIE 2303*, A. F. Laine and M. A. Unser, Eds., 1994, pp. 2–14.

[28] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis.* Cambridge University Press, 2000.

[29] L. Atlas, J. Droppo, and J. McLaughlin, "Optimizing time-frequency distributions for automatic classification," *SPIE*, vol. 3162, pp. 161–171, 1997.

[30] B. Gillespie and L. Atlas, "Optimizing time-frequency kernels for classification," *IEEE Transactions of Signal Processing*, vol. 49, pp. 1341–4, 2001.

[31] M. Davy and C. Doncarli, "Optimal kernels of time-frequency representations for signal classification," in *IEEE International Symposium on Time-Frequency and Time Scale*, 1998, pp. 581–584.

[32] M. Davy, C. Doncarli, and G. F. Boudreaux-Bartles, "Improved optimization of time frequency-based signal classifiers," *IEEE Signal Processing Letters*, vol. 8, pp. 52–57, 2001.

[33] M. Davy, A. Gretton, A. Doucet, and P. J. W. Rayner, "Optimized support vector machines for nonstationary signal classification," *IEEE Signal Processing Letters*, vol. 9, pp. 442–5, 2002.

[34] R. G. Baraniuk and D. L. Jones, "A signal-dependent time-frequency representation: Optimal kernel design," *IEEE Transactions on Signal Processing*, vol. 41, pp. 1589–1601, 1993.

[35] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[36] V. Vapnik, *The Nature of Statistical Learning Theory.* New York: Springer-Verlag, 1995.

[37] M. E. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems*, 2000.

[38] ——, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning*, vol. 1, pp. 211–244, 2001.

[39] M. R. Gupta, L. Cazzanti, and A. Koppal, "Maximum entropy generative models for similarity-based learning," in *IEEE International Symposium on Information Theory*, 2007.

[40] S. Hochreiter, M. C. Mozer, and K. Obermayer, "Coulomb classifiers: Generalizing support vector machines via an analogy to electrostatic systems," in *Advances in Neural Information Processing Systems 15*, 2003, p. 545552.

[41] S. Hochreiter and K. Obermayer, "Support vector machines for dyadic data," *Neural Computation*, vol. 18, no. 6, pp. 14721510,, 2006.

[42] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[43] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Neural Information Processing Systems*, 2001.

[44] O. Chapelle, J. Weston, and B. Schölkopf, "Cluster kernels for semi-supervised learning," in *Neural Information Processing Systems*, 2003.

[45] X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty, "Nonparametric transforms of graph kernels for semi-supervised learning," in *Neural Information Processing Systems*, 2005.

[46] M. Zampolli, A. Tesei, G. Canepa, F. Jensen, J. Fawcett, and S. Philips, "Numerical modeling of scattering from completely and partially filled elastic targets for the detection and classification of submerged objects," in *European Conference on Underwater Acoustics*, 2006.

[47] P. W. Cooper and S. R. Kurowski, *Introduction to the Technology of Explosives.* New York: Wiley-VCH, 1996.

[48] F. W. Young, C. H. Null, W. Sarle, and D. L. Hoffman, *Proximity and preference: Problems in the multi-dimensional analysis of large data sets.* Minneapolis, MN: University of Minnesota Press., 1981, ch. Interactively ordering the similarities among a large set of stimuli.

[49] V. Young, "Application of musical timbre discrimination features to active sonar classification," Master's thesis, Dalhousie University, Halifax, NS, 2005.

[50] B. Schölkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[51] R. L. Goldstone and A. W. Kersten, *Comprehensive Handbook of Psychology.* New Jersey: Wiley, 2003, vol. 4, ch. Concepts and Categorization, pp. 599–621.

[52] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, 2001.

[53] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* Wiley-Interscience, 2001.

[54] J. Kliewer and A. Mertins, "Au:lio subband coding with improved representation of transient signal segments," in *European Signal Processing Conference*, 1998.

Appendix A

Standard Features

Below are a list of standard features that have been identified in the published literature [1], [14], [49]. These features are used in Chapter 5 to assess their perceptual relevance. The results of this perceptual analysis is summarized in the tables following the feature definitions.

Time Features:

$x(t)$	=	recorded time-series
$v(t)$	=	hilbert envelope of the recorded time-series
t_0	=	time at the start of the signal of interest as defined in [54]
t_1	=	time at the end of the signal of interest as defined in [54]
t_p	=	time at which the peak of the signal occurs
N_0	=	noise floor energy

rise time -

$$\phi_1 = t_p - t_0 \tag{A.1}$$

fall time -

$$\phi_2 = t_1 - t_p \tag{A.2}$$

duration -

$$\phi_3 = t_1 - t_0 \tag{A.3}$$

centroid -

$$\phi_4 = \frac{\sum_t t \cdot v(t)}{\sum_t v(t)} \tag{A.4}$$

(peak Magnitude)/(avg rms value) -

$$\phi_5 = \frac{t_p}{N_0} \tag{A.5}$$

Spectral Features:

$X(\omega)$	=	power spectral density function
ω_N	=	Nyquist frequency

specCentroid -

$$\phi_6 = \frac{\sum_\omega \omega \cdot X(\omega)}{\sum_\omega X(\omega)} \tag{A.6}$$

specSpread -

$$\phi_7 = \frac{\sum_\omega (\omega - \phi_6)^2 \cdot X(\omega)}{\sum_\omega X(\omega)} \tag{A.7}$$

specFlatness -

$$\phi_8 = \frac{\sqrt[N]{\prod_\omega X(\omega)}}{\frac{1}{N}\sum_\omega X(\omega)} \tag{A.8}$$

specRolloff -

$$\phi_9 = \underset{\omega_R}{\operatorname{argmax}} \sum_{\omega=0}^{\omega_R} X(\omega) \leq .85 \sum_{\omega=0}^{\omega_N} X(\omega) \tag{A.9}$$

Time-Frequency Features:

$S(t,\omega)$ = Spectrogram

$t_{\omega_1,0}$ = time at the start of the signal in subband ω_1

$t_{\omega_1,p}$ = time at which the peak of the signal in subband ω_1

specFlux -

$$\phi_{10} = \max_t \sum_\omega S(t,\omega) - S(t-1,\omega) \tag{A.10}$$

maxSBCorr -

$$\phi_{11} = \max_{\omega_1,\ \omega_2} \frac{\sum_t S(t,\omega_1) \cdot S(t,\omega_2)}{(\sum_t S(t,\omega_1))(\sum_t S(t,\omega_2))} \tag{A.11}$$

maxSBCorrFreq -

$$\phi_{12} = \underset{\omega_1}{\operatorname{argmax}} \sum_{\omega_2} \frac{\sum_t S(t,\omega_1) \cdot S(t,\omega_2)}{(\sum_t S(t,\omega_1))(\sum_t S(t,\omega_2))} \tag{A.12}$$

minSBCorr -

$$\phi_{13} = \min_{\omega_1,\ \omega_2} \frac{\sum_t S(t,\omega_1) \cdot S(t,\omega_2)}{(\sum_t S(t,\omega_1))(\sum_t S(t,\omega_2))} \tag{A.13}$$

minSBCorrFreq -

$$\phi_{14} = \underset{\omega_1}{\operatorname{argmin}} \sum_{\omega_2} \frac{\sum_t S(t,\omega_1) \cdot S(t,\omega_2)}{(\sum_t S(t,\omega_1))(\sum_t S(t,\omega_2))} \tag{A.14}$$

MaxGSBAttackSlope -

$$\phi_{15} = \max_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_0,\omega_1)}{t_{\omega_1,p} - t_0} \tag{A.15}$$

MaxGSBAttackSlopeFreq -

$$\phi_{16} = \underset{\omega_1}{\operatorname{argmax}} \frac{S(t_{\omega_1,p},\omega_1) - S(t_0,\omega_1)}{t_{\omega_1,p} - t_0} \tag{A.16}$$

MaxGSBAttackTime -

$$\phi_{17} = \max_{\omega_1} t_{\omega_1,p} - t_0 \tag{A.17}$$

MaxGSBAttackTimeFreq -

$$\phi_{18} = \underset{\omega_1}{\operatorname{argmax}}\, t_{\omega_1,p} - t_0 \tag{A.18}$$

MeanGSBAttackSlope -

$$\phi_{19} = \frac{1}{N}\sum_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_0,\omega_1)}{t_{\omega_1,p} - t_0} \tag{A.19}$$

MinGSBAttackSlope -

$$\phi_{20} = \min_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_0,\omega_1)}{t_{\omega_1,p} - t_0} \tag{A.20}$$

MinGSBAttackSlopeFreq -

$$\phi_{21} = \operatorname*{argmin}_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_0,\omega_1)}{t_{\omega_1,p} - t_0} \tag{A.21}$$

MinGSBAttackTime -

$$\phi_{22} = \min_{\omega_1} t_{\omega_1,p} - t_0 \tag{A.22}$$

MinGSBAttackTimeFreq -

$$\phi_{23} = \operatorname*{argmin}_{\omega_1} t_{\omega_1,p} - t_0 \tag{A.23}$$

MaxLSBAttackSlope -

$$\phi_{24} = \max_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_{\omega_1,0},\omega_1)}{t_{\omega_1,p} - t_{\omega_1,0}} \tag{A.24}$$

MaxLSBAttackSlopeFreq -

$$\phi_{25} = \operatorname*{argmax}_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_{\omega_1,0},\omega_1)}{t_{\omega_1,p} - t_{\omega_1,0}} \tag{A.25}$$

MaxLSBAttackTime -

$$\phi_{26} = \max_{\omega_1} t_{\omega_1,p} - t_{\omega_1,0} \tag{A.26}$$

MaxLSBAttackTimeFreq -

$$\phi_{27} = \operatorname*{argmax}_{\omega_1} t_{\omega_1,p} - t_{\omega_1,0} \tag{A.27}$$

MeanLSBAttackSlope -

$$\phi_{28} = \frac{1}{N}\sum_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_{\omega_1,0},\omega_1)}{t_{\omega_1,p} - t_{\omega_1,0}} \tag{A.28}$$

MinLSBAttackSlope -

$$\phi_{29} = \min_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_{\omega_1,0},\omega_1)}{t_{\omega_1,p} - t_{\omega_1,0}} \tag{A.29}$$

MinLSBAttackSlopeFreq -

$$\phi_{30} = \operatorname*{argmin}_{\omega_1} \frac{S(t_{\omega_1,p},\omega_1) - S(t_{\omega_1,0},\omega_1)}{t_{\omega_1,p} - t_{\omega_1,0}} \tag{A.30}$$

MinLSBAttackTime -

$$\phi_{31} = \min_{\omega_1} t_{\omega_1,p} - t_{\omega_1,0} \tag{A.31}$$

MinLSBAttackTimeFreq -

$$\phi_{32} = \operatorname*{argmin}_{\omega_1} t_{\omega_1,p} - t_{\omega_1,0} \tag{A.32}$$

localSBAttackDiff -

$$\phi_{33} = \frac{1}{N^2}\sum_{\omega_1,\omega_2} t_{\omega_1,0} - t_{\omega_2,0} \tag{A.33}$$

Table A.1: Correlation values of the standard features against U.S. Navy MDS dimensions one and two

Feature	**Correlation to Dimension 1**	**Correlation to Dimension 2**
rise time	-0.2380	0.1943
fall time	-0.5182	0.2654
duration	-0.4809	0.2861
centroid	0.0886	-0.2401
(peak Magnitude)/(avg rms value)	0.3199	-0.1324
specCentroid	-0.0017	-0.2968
specSpread	-0.2024	0.0037
specFlatness	-0.0809	0.0420
specFlux	0.3001	-0.2080
specRolloff	-0.3008	-0.2892
maxSBCorr	0.0787	0.1360
maxSBCorrFreq	0.6068	0.2372
minSBCorr	0.1056	-0.0553
minSBCorrFreq	-0.4276	-0.1090
MaxGSBAttackSlope	0.1914	-0.1321
MaxGSBAttackSlopeFreq	0.3834	0.2475
MaxGSBAttackTime	0.1407	-0.0788
MaxGSBAttackTimeFreq	-0.2321	-0.3831
MeanGSBAttackSlope	0.2917	-0.1527
MinGSBAttackSlope	-0.1482	-0.0434
MinGSBAttackSlopeFreq	-0.4733	-0.2728
MinGSBAttackTime	-0.1361	0.0712
MinGSBAttackTimeFreq	0.2613	0.3969
MaxLSBAttackSlope	0.1300	-0.1470
MaxLSBAttackSlopeFreq	0.3941	0.2727
MaxLSBAttackTime	0.0241	-0.0825
MaxLSBAttackTimeFreq	-0.2932	-0.3434
MeanLSBAttackSlope	0.3193	-0.3378
MinLSBAttackSlope	0.1461	-0.0265
MinLSBAttackSlopeFreq	-0.2043	-0.2225
MinLSBAttackTime	0.1427	-0.0237
MinLSBAttackTimeFreq	0.1286	-0.1817
localSBAttackDiff - mean	-0.0791	-0.2933

Table A.2: Correlation values of the standard features against Boundary 2004 MDS dimensions one and two

Feature	Correlation to Dimension 1	Correlation to Dimension 2
rise time	0.5029	0.5449
fall time	0.1331	0.4799
duration	0.1428	0.3522
centroid	0.3206	-0.1928
(peak Magnitude)/(avg rms value)	-0.5723	-0.1637
specCentroid	-0.3776	-0.3261
specSpread	0.0855	0.1534
specFlatness	-0.5842	0.1953
specFlux	-0.5816	-0.3105
specRolloff	-0.8205	0.0278
maxSBCorr	-0.5878	0.0657
maxSBCorrFreq	0.4625	-0.4360
minSBCorr	-0.6469	-0.0364
minSBCorrFreq	-0.1918	0.3063
MaxGSBAttackSlope	-0.2932	-0.1931
MaxGSBAttackSlopeFreq	0.6098	-0.3353
MaxGSBAttackTime	0.2750	-0.0634
MaxGSBAttackTimeFreq	-0.3958	0.0387
MeanGSBAttackSlope	-0.5845	-0.3340
MinGSBAttackSlope	-0.4786	-0.2675
MinGSBAttackSlopeFreq	-0.5280	0.1435
MinGSBAttackTime	-0.1603	-0.0846
MinGSBAttackTimeFreq	0.2644	-0.4014
MaxLSBAttackSlope	-0.2524	-0.3696
MaxLSBAttackSlopeFreq	0.5526	-0.3213
MaxLSBAttackTime	0.4037	0.0797
MaxLSBAttackTimeFreq	-0.2265	0.3122
MeanLSBAttackSlope	-0.6684	-0.3797
MinLSBAttackSlope	-0.6254	-0.2035
MinLSBAttackSlopeFreq	-0.3399	0.1455
MinLSBAttackTime	-0.1052	0.0018
MinLSBAttackTimeFreq	0.2394	-0.0518
localSBAttackDiff - mean	0.7016	-0.2164

Table A.3: Correlation values of the standard features against Modeled Mine parameter dimensions one and two

Feature	Correlation to Dimension 1	Correlation to Dimension 2
rise time	0.6937	-0.0331
fall time	-0.6385	-0.2930
duration	-0.7316	-0.3172
centroid	0.0846	-0.0872
(peak Magnitude)/(avg rms value)	0.2452	-0.2449
specCentroid	-0.8088	-0.0763
specSpread	0.8521	0.0570
specFlatness	0.7367	0.1385
specFlux	0.7384	-0.0478
specRolloff	-0.1539	-0.0208
maxSBCorr	-0.6615	-0.0788
maxSBCorrFreq	-0.5793	0.0795
minSBCorr	0.2082	-0.1707
minSBCorrFreq	0.5627	0.1166
MaxGSBAttackSlope	0.7043	-0.0192
MaxGSBAttackSlopeFreq	0.0540	-0.0300
MaxGSBAttackTime	-0.1658	-0.1568
MaxGSBAttackTimeFreq	0.5355	0.0126
MeanGSBAttackSlope	0.7022	-0.1184
MinGSBAttackSlope	-0.2502	-0.1515
MinGSBAttackSlopeFreq	0.5974	0.0962
MinGSBAttackTime	-0.7725	-0.1967
MinGSBAttackTimeFreq	-0.3466	-0.2261
MaxLSBAttackSlope	0.7292	-0.0875
MaxLSBAttackSlopeFreq	0.4218	0.1309
MaxLSBAttackTime	0.2744	0.1031
MaxLSBAttackTimeFreq	0.2744	0.0029
MeanLSBAttackSlope	0.2658	-0.0813
MinLSBAttackSlope	-0.6024	0.0343
MinLSBAttackSlopeFreq	0.1031	-0.3699
MinLSBAttackTime	-0.6496	-0.1932
MinLSBAttackTimeFreq	0.0814	-0.3152
localSBAttackDiff - mean	0.6235	0.0982